Entropy of Hidden Markov Processes via Cycle Expansion.

Armen E. Allahverdyan
Yerevan Physics Institute, Alikhanian Brothers Street 2, Yerevan 375036, Armenia

(Dated: October 23, 2008)

Hidden Markov Processes (HMP) is one of the basic tools of the modern probabilistic modeling. The characterization of their entropy remains however an open problem. Here the entropy of HMP is calculated via the cycle expansion of the zeta-function, a method adopted from the theory of dynamical systems. For a class of HMP this method produces exact results both for the entropy and the moment-generating function. The latter allows to estimate, via the Chernoff bound, the probabilities of large deviations for the HMP. More generally, the method offers a representation of the moment-generating function and of the entropy via convergent series.

PACS numbers: 89.70.Cf, 05.20.-y

I. INTRODUCTION.

Hidden Markov Processes (HMP) are generated by a Markov process observed via a memory-less noisy channel. They are widely employed in various areas of probabilistic modeling [1, 2, 3, 4]: information theory, signal processing, bioinformatics, mathematical economics, linguistics, *etc.* One of the main reasons for these numerous applications is that HMP present simple and flexible models for a history-dependent random process. This is in contrast to the Markov process, where the history is irrelevant, since the future of the process depends on its present state only.

Much attention was devoted to the entropy of HMP [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. It characterizes the information content (minimal number of bits needed for a reliable encoding) of HMP viewed as a probabilistic source of information. More specifically, the realizations generated in the long run of a random ergodic process, e.g. HMP, are divided into two sets [6, 8]. The first (typical) set is the smallest set of realizations with the overall probability close to one. The rest of realizations are contained in the second, low-probability set. Now the entropy characterizes the number of elements in the typical set [6, 8]. When HMP is employed as a model for information transmission over a noisy channel, the entropy is still important, since it is the basic non-trivial component of the channel capacity (other components needed for reconstructing the channel capacity are normally easier to calculate and characterize).

However, there is no direct formula for the entropy of HMP, in contrast to the Markov case where such a formula is well-known [5, 6, 8]. Thus people studied the entropy via expansions around various limiting cases, or via upper and lower bounds [6, 10, 11, 12, 13, 14, 15, 16, 17]. There is also a general formalism that expresses the entropy of HMP via the solution of an integral equation [7, 8, 9]. This formalism is however relatively difficult to apply in practice.

Once the entropy characterizes the number of typical long-run realizations, it is of interest to estimate the probability of atypical realizations. These estimates are standardly given via the moment-generating function of the random process [6, 8]. The knowledge of this function also allows to reconstruct the entropy [6, 8].

This paper presents a method for calculating the moment-generating function of HMP. The method is adopted from the theory of chaotic dynamical systems, where it is known as the cycle expansion of the zeta-function [25, 27]. We show that in a certain class of HMP one can obtain exact expressions for the moment-generating function and for the entropy. For other cases the method offers analytic approximations of the moment-generating function via convergent power series.

We attempted to make this paper self-contained and organized it as follows. Section II defines the HMP, settles some notations, and recalls how to express the probabilities of HMP via a random matrix product. In section III we briefly review the main facts about the entropy of an ergodic process and the corresponding typical (highly probable) set of realizations. The main purpose of section IV is to relate the entropy of HMP to the spectral radius of the corresponding random matrix product. This is done via the Lyapunov exponent of the random matrix product. Section V discusses the moment-generating function of HMP. This function is employed (via Chernoff bounds) for characterizing the atypical (improbable) realizations of HMP. Section VI shows how to calculate the entropy and the generating function via the zeta-function and the periodic orbit expansion. Section VII discusses one of the simplest examples of HMP and presents exact expressions for its entropy and the moment-generating function. Here we also apply the moment-generating function for estimating atypical realizations of the HMP. Section VIII studies another popular model for HMP, binary symmetric HMP. It is shown that the presented approach reproduces known approximate results and predicts several new ones. The last section shortly summarizes the obtained results. Some issues, which are either too technical or too general for the present purposes, are discussed in Appendices.

II. DEFINITION OF THE HIDDEN MARKOV PROCESS.

In this section we recall the definition of the Hidden Markov Process (HMP); see [1, 2] for reviews.

Let a discrete-time random process $\mathcal{S} = \{\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2, ...\}$ be Markovian, with time-independent conditional probability

$$\Pr[\mathcal{S}_k = s_k|\mathcal{S}_{k-1} = s_{k-1}] = \Pr[\mathcal{S}_{k+l} = s_k|\mathcal{S}_{k-1+l} = s_{k-1}] = p(s_k|s_{k-1}), \tag{1}$$

where l is an integer. Each realization s of the random variable $\mathcal{S}$ takes values $s = 1, ..., L$. The joint probability of the Markov process reads

$$\Pr[\mathcal{S}_N = s_N, ..., \mathcal{S}_0 = s_0] = p(s_N|s_{N-1}) \dots p(s_1|s_0)p(s_0) = \prod_{k=N}^{1} p(s_k|s_{k-1})\, p(s_0), \tag{2}$$

where $p(s_0)$ is the initial probability. The conditional probabilities $p(s_k|s_{k-1})$ define the $L \times L$ transition matrix $\mathbb{P}$:

$$\mathbb{P}_{s_k\, s_{k-1}} = p(s_k|s_{k-1}). \tag{3}$$

We assume that the Markov process $\mathcal{S}$ is mixing [18]: it has a unique stationary distribution $p_{\rm st}(s)$,

$$\sum_{s'=1}^{L} p(s|s')p_{\rm st}(s') = p_{\rm st}(s), \tag{4}$$

that is established from any initial probability in the long time limit. The transition matrix $\mathbb{P}$ has always one eigenvalue equal to 1 [since $\mathbb{P}$ has a left eigenvector $(1, ..., 1)$], and the modules [absolute values] of all other eigenvalues are not larger than one [1]. The mixing feature however demands that the eigenvalue equal to 1 is non-degenerate and the modules of all other eigenvalues are smaller than 1 [18]. A sufficient condition for mixing is that all the conditional probabilities $p(s_{i+1}|s_i)$ of the Markov process are positive [18] [2]. Taking $p(s) = p_{\rm st}(s)$ in (2) makes the process $\mathcal{S}$ stationary.

Let random variables $\mathcal{X}_i$, with realizations $x_i = 1, .., M$, be noisy observations of $\mathcal{S}_i$: the (time-invariant) conditional probability of observing $\mathcal{X}_i = x_i$ given the realization $\mathcal{S}_i = s_i$ of the Markov process is $\pi(x_k|s_k)$. The joint probability of the original process and its noisy observations reads

$$\begin{aligned} P(s_N, \dots, s_0; x_N, \dots, x_1) &= \prod_{k=N}^{1} \pi(x_k|s_k)p(s_k|s_{k-1})p_{\rm st}(s_0) && (5)\\ &= T_{s_N\, s_{N-1}}(x_N)...T_{s_1\, s_0}(x_1)\, p_{\rm st}(s_0), && (6)\end{aligned}$$

where the $L \times L$ transfer-matrix $T(x)$ with matrix elements $T_{s_i\, s_{i-1}}(x)$ is defined as

$$T_{s_i\, s_{i-1}}(x) = \pi(x|s_i)\, p(s_i|s_{i-1}). \tag{7}$$

Thus $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, ...\}$, called hidden Markov process, results from observing the Markov process $\mathcal{S}$ through a memory-less process with the conditional probability $\pi(x|s)$. The composite process $\mathcal{S}\mathcal{X}$ is Markovian as well.

The probabilities for the process $\mathcal{X}$ are represented via the transfer matrix product (similar representation were employed in [11, 12])

$$\begin{aligned} P(\mathbf{x}_{N\dots1}) &= \langle {\rm un}|\mathbb{T}(\mathbf{x}_{N\dots1})|{\rm st}\rangle, && (8)\\ \mathbb{T}(\mathbf{x}_{N\dots1}) &\equiv \prod_{k=N}^{1} T(x_k), && (9)\\ \mathbf{x}_{N\dots1} &\equiv (x_N, ..., x_1), && (10)\end{aligned}$$

[1] Indeed, $\sum_k \mathbb{P}_{ik}x_k = \nu x_i$ implies $|\sum_k \mathbb{P}_{ik}x_k| \le \sum_k \mathbb{P}_{ik}|x_k| = |\nu||x_i|$, which then leads to $|\nu| \le 1$.

[2] Weaker sufficient conditions for mixing are that *i)* for any (i, j) there exists a positive integer m_{ij} such that $(\mathbb{P}^{m_{ij}})_{ij} > 0$, i.e., for some power of the matrix its entries are positive, and *ii)* $\mathbb{P}$ has at least one positive diagonal element [18]. If we do assume the first condition, but do not assume the second one, the eigenvalue 1 of $\mathbb{P}$ is [algebraically and thus geometrically] non-degenerate, and is not smaller than the absolute values of all other eigenvalues [18]. The corresponding [unique] eigenvector has strictly positive components. However, it may be that the module of some other eigenvalue(s) is equal to 1 thus preventing the proper mixing, but still allowing for ergodicity due to condition *i)*.

where we used the bra(c)ket notations: $|\mathrm{st}\rangle$ is the column vector with elements $p_{\mathrm{st}}(k)$, $k=1,...,L$, and $\langle\mathrm{un}| = (1,...,1)$.

The HMP defined by (8) is (in general) not a Markov process, i.e., its probabilities do not factorize as in (2). Thus the history of the process can become relevant. This is the underlying reason for widespread applications of HMP.

The process $\mathcal{X}$ is stationary due to the stationarity of $\mathcal{S}$:

$$\Pr[\mathcal{X}_{N+l} = x_N, ..., \mathcal{X}_{l+1} = x_1] = \Pr[\mathcal{X}_N = x_N, ..., \mathcal{X}_1 = x_1] = P(x_N, ..., x_1), \tag{11}$$

where l is a positive integer.

In addition, $\mathcal{X}$ inherits the mixing feature from the underlying Markov process $\mathcal{S}$ [2], because the observation process by itself is memoryless: $\pi(x_k|s_k) = \pi(x_k|s_k, s_{k-1}, ..., s_0)$. (The general definitions of ergodicity and mixing are reminded below.)

A. Notations for the eigenvalues and singular values.

For future purposes we concretise some notations. For a matrix A, let $l_0[A], l_1[A], ...$ be the modules of its eigenvalues. We order $l_k[A]$ as

$$\lambda[A] \equiv l_0[A] \,\geq\, l_1[A] \,\geq \ldots, \tag{12}$$

$\lambda[A]$ is called the spectral radius of A [18]. If A has non-negative matrix elements, the spectral radius is an eigenvalue by itself [18]. Here are two obvious features of the function λ (d is a positive integer):

$$\lambda[A^d] = (\lambda[A])^d, \tag{13}$$

$$\lambda[AB] = \lambda[BA], \tag{14}$$

where (14) follows from the fact that AB and BA have identical eigenvalues: $AB|\psi\rangle = \nu|\psi\rangle$ implies $BA\ (B|\psi\rangle) = \nu B|\psi\rangle$.

Let $A^\dagger$ be the complex conjugate of A. The singular values $\sigma_k[A] \geq 0$ for a matrix A are the eigenvalues of a hermitean matrix $\sqrt{AA^\dagger}$ or, equivalently, of $\sqrt{A^\dagger A}$; see Appendix A for a brief reminder on the features of the singular values. We order $\sigma_k[A]$ as

$$\sigma_0[A] \geq \sigma_1[A] \geq .... \tag{15}$$

III. ENTROPY AND TYPICAL SET OF ERGODIC PROCESSES.

The N-block entropy of a stationary [not necessarily Hidden Markov] random process $\mathcal{X}$ is defined as [5, 6, 8]

$$H(N) = H(\mathcal{X}_1, ..., \mathcal{X}_N) \equiv -\sum_{\mathbf{x}_{N...1}} P(\mathbf{x}_{N...1}) \ln P(\mathbf{x}_{N...1}), \tag{16}$$

where the probability $P(\mathbf{x}_{N...1})$ is given as in (8), and where $\mathbf{x}_{N...1}$ is defined in (10). Various features of $H(N)$ and of several related quantities are discussed in Appendix B.

Using (16) one now defines the entropy (rate) of the random process $\mathcal{X}$ as [5, 6, 8]

$$h = \lim{}_{N\to\infty} \frac{H(N)}{N}. \tag{17}$$

Alternative representations of h are recalled in Appendix B. In particular, h is the uncertainty [per unit of time] of the random process given its long history.

For ergodic processes the above definition of entropy can be related to a single, long sequence of realizations [5, 6, 8]. First of all let us recall that the process $\mathcal{X}$ is ergodic if it satisfies to the weak law of large numbers (time average is equal to the space average): for any function f with a finite expectation value $\bar{f} \equiv \sum_{x_k,...,x_0} f[x_k, ..., x_0] P(x_k, ..., x_0)$, we have probability-one convergence for $N \to \infty$ [5, 6, 8]:

$$\frac{1}{N} \sum_{n=0}^{N-1} f[\mathcal{X}_{n+k}, ..., \mathcal{X}_n] \to \bar{f}, \tag{18}$$

i.e., for any positive numbers ε and δ, there is such an integer $\mathcal{N}(\varepsilon,\delta)$ that for all $N > \mathcal{N}(\varepsilon,\delta)$,

$$\Pr\left[\left|\frac{1}{N}\sum_{n=0}^{N-1} f[\mathcal{X}_{n+k},...,\mathcal{X}_n] - \bar{f}\right| \geq \varepsilon\right] \leq \delta. \tag{19}$$

Several alternative definitions of ergodicity are discussed in [35][3].

Now the McMillan lemma states that for an ergodic process the entropy (17) characterizes individual realizations in the sense of probability-one convergence for $N \to \infty$ [5, 6, 8] [4]:

$$-\frac{1}{N}\ln P(\mathbf{x}_{N\dots 1}) \to h \quad \text{or} \quad \Pr\left[\left|-\frac{1}{N}\ln P(\mathbf{x}_{N\dots 1}) - h\right| \leq \varepsilon\right] \geq 1-\delta. \tag{20}$$

Based on (20) one defines the typical set $\Omega_N^*(\varepsilon)$ as the set of all $\mathbf{x}_{N\dots 1}$, which satisfy to

$$h - \varepsilon \leq -\frac{1}{N}\ln P(\mathbf{x}_{N\dots 1}) \leq h + \varepsilon. \tag{21}$$

Now (20) implies that $\Pr[\mathbf{x}_{N\dots 1} \in \Omega_N^*(\varepsilon)] \geq 1-\delta$, i.e., the overall probability of $\Omega_N^*(\varepsilon)$ converges to one in the limit $N \to \infty$. Since all elements in $\Omega_N^*(\varepsilon)$ have approximately equal probabilities, the number of elements $|\Omega_N^*(\varepsilon)|$ in $\Omega_N^*(\varepsilon)$ scales as e^{Nh}. More precisely, this number is estimated from (20, 21) as [5]

$$(1-\delta)e^{N(h-\varepsilon)} \leq |\Omega_N^*(\varepsilon)| \leq e^{N(h+\varepsilon)}. \tag{22}$$

Relations similar to (21) will be frequently written as

$$P(\mathbf{x}_{N\dots 1}) \simeq e^{-Nh} \quad \text{for} \quad \mathbf{x}_{N\dots 1} \in \Omega_N^*, \tag{23}$$

meaning that the precise sense of the asymptotic relation $\simeq$ for $N \to \infty$ can be clarified upon introducing proper ϵ and δ.

IV. LYAPUNOV EXPONENTS AND ENTROPY.

The purpose of this section is to establish relation (29) between the entropy of a Hidden Markov Process, and the spectral radius of the associated random matrix product (8). The reader may skip this section, if this relation is taken granted.

A. Singular values of the random-matrix product.

The actual calculation of the entropy h for non-Markov processes meets (in general) considerable difficulties. (For Markov processes definition (17) applies directly leading to the well-known formula for the entropy [5].) The first step in calculating the entropy h for a Hidden Markov Process (HMP) is to relate h to the large-N behaviour of the $L \times L$ matrix $\mathbb{T}(\mathbf{x}_{N\dots 1})$, which defines the probability of HMP; see (8, 9). Recall that $\mathbb{T}(\mathbf{x}_{N\dots 1})$ is a function of the random process $\mathcal{X}$. Assume that *i)* $\mathcal{X}$ is stationary, as is the case after (11). *ii)* The average logarithm of the maximal singular value of $T(x)$ is finite: $\langle \ln \sigma_0[T(x)]\rangle < \infty$. *iii)* $\mathcal{X}$ is ergodic. Then the subadditive ergodic theorem applies claiming for $N \to \infty$ the probability-one convergence [19, 20]:

$$-\frac{1}{N}\ln \sigma_k[\mathbb{T}(\mathbf{x}_{N\dots 1})] \to \mu_k, \quad k = 0, \dots, L-1, \tag{24}$$

[3] One such definition is worth mentioning: $\mathcal{X}$ is ergodic if for any k, m and s: $\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\Pr[\mathcal{X}_{n+k} = x_k, ..., \mathcal{X}_n = x_0, \mathcal{X}_{m+s} = y_m, ..., \mathcal{X}_s = y_0] = P(x_k,...,x_0)P(y_m,...,y_0)$. This definition admits a straightforward and important generalization. $\mathcal{X}$ is called mixing if the above relation holds without the time-averaging $\frac{1}{N}\sum_{n=0}^{N-1}$, but in the limit $n \to \infty$.

[4] The McMillan lemma contains two essential steps [5]. First is to realize that although the definition (18) of ergodicity does not apply directly to $\frac{1}{N}\ln P(\mathbf{x}_{N\dots 1})$, it does apply to the probability $Q_m(\mathbf{x}_{N\dots 1}) = P(x_1,...,x_m)\prod_{i=1}^{N-m} P(x_{m+i}|x_{m+i-1},...,x_i)$, which defines an approximation of the original ergodic process by a m-order Markov process. In the second step using a chain of inequalities $\Pr[|\ln x| \geq n\varepsilon] \leq \frac{1}{n\varepsilon}\overline{|\ln x|} \leq \frac{1}{n\varepsilon}\overline{(2x - \ln x)}$, one proves that for any stationary [not necessarily ergodic] process $Q_m(\mathbf{x}_{N\dots 1})$ is indeed a good approximation in the sense of $\frac{1}{N}\ln\frac{Q_m(\mathbf{x}_{N\dots 1})}{P(\mathbf{x}_{N\dots 1})} \simeq 0$ for $N \gg m \to \infty$.

where $\sigma_k[\mathbb{T}(\mathbf{x}_{N\dots1})]$ are the singular values of $\mathbb{T}(\mathbf{x}_{N\dots1})$ (see section II A for notations), and where μ_k are called Lyapunov exponents. According to (15) they are ordered as $\mu_0 \leq \mu_1 \leq \dots$.

Using the definition (21) of the typical set, (24) can be written as an asymptotic relation $\sigma_k[\mathbb{T}(\mathbf{x}_{N\dots1})] \simeq e^{-N\mu_k}$ for $\mathbf{x}_{N\dots1} \in \Omega_N$ and sufficiently large N [21]. Moreover, employing the singular value decomposition [see Appendix A], one represents $\mathbb{T}(\mathbf{x}_{N\dots1})$ for $N \to \infty$ and $\mathbf{x}_{N\dots1} \in \Omega_N^*$ as

$$\mathbb{T}(\mathbf{x}_{N\dots1}) \simeq \mathrm{diag}\left[e^{-N\mu_0}, \dots, e^{-N\mu_{L-1}}\right] U(\mathbf{x}), \tag{25}$$

where diag $[a, \dots, b]$ is a diagonal matrix with entries $a, \dots, b$, and where $U(\mathbf{x})$ is an orthogonal matrix. The fact that (for $N \to \infty$) the matrix U does not depend on N (but does in general depend on the realization $\mathbf{x}$) is a consequence of the Oseledec theorem [21, 22].

Thus the meaning of (25) is that the essential dependence of $\mathbb{T}(\mathbf{x}_{N\dots1})$ on N is contained in the singular values $e^{-N\mu_k}$, while $U(\mathbf{x})$ does not depend on N for $N \to \infty$.

B. Eigenvalues of the random-matrix product.

The above reasoning by itself is silent about the eigenvalues of $\mathbb{T}(\mathbf{x}_{N\dots1})$. Since the matrix $\mathbb{T}(\mathbf{x}_{N\dots1})$ is in general not normal, i.e., the commutator of $\mathbb{T}(\mathbf{x}_{N\dots1})$ with its transpose $\mathbb{T}^\dagger(\mathbf{x}_{N\dots1})$ is not zero, the modules $l_k[\mathbb{T}(\mathbf{x}_{N\dots1})]$ of its eigenvalues are not automatically equal to its singular values $e^{-N\mu_k}$; see Appendix A. For us the knowledge of the spectral radius $\lambda[\mathbb{T}(\mathbf{x}_{N\dots1})]$ will be important, because for calculating the entropy we shall employ a method that essentially relies on the features (13, 14), which hold for the eigenvalues, but do not hold for singular values.

It is shown in Appendix D that the representation (25) can be used for deducing that in the limit $N \to \infty$ and for $\mathbf{x}_{N\dots1} \in \Omega_N^*$ the spectral radius $\lambda[\mathbb{T}(\mathbf{x}_{N\dots1})]$ of $\mathbb{T}(\mathbf{x}_{N\dots1})$ behaves as [recall (12)]

$$\lambda[\mathbb{T}(\mathbf{x}_{N\dots1})] \simeq e^{-N\mu_0}, \tag{26}$$

where μ_0 is the so called top Lyapunov exponent. Appendix D discusses under which *generic* conditions (26) holds; see also [23] in this context.

Using (8) we have asymptotically for $N \to \infty$ and $\mathbf{x}_{N\dots1} \in \Omega_N^*$

$$\mathbb{T}(\mathbf{x}_{N\dots1}) \simeq e^{-N\mu_0} |\, R(\mathbf{x})\, \rangle\langle\, L(\mathbf{x})\, | + \mathcal{O}[e^{-N\nu_1(\mathbf{x}_{N\dots1})}], \tag{27}$$

$$P(\mathbf{x}_{N\dots1}) \simeq e^{-N\mu_0 + \mathcal{O}(1)} + \mathcal{O}[e^{-N\nu_1(\mathbf{x}_{N\dots1})}], \tag{28}$$

where we denoted $l_1[\mathbb{T}(\mathbf{x}_{N\dots1})] \equiv e^{-N\nu_1(\mathbf{x}_{N\dots1})}$ [see (12)], and where $|\, R(\mathbf{x})\, \rangle$ and $|\, L(\mathbf{x})\, \rangle$ are, respectively, the right and left eigenvectors of $\mathbb{T}(\mathbf{x}_{N\dots1})$; see Appendix A. They do not depend on N (for $N \to \infty$) for the same reason as U in (25) does not depend on N. In writing down (27) we assumed that the spectral radius $\lambda[\mathbb{T}(\mathbf{x}_{N\dots1})]$ is not a degenerate eigenvalue of $\mathbb{T}(\mathbf{x}_{N\dots1})$, or at least that its algebraic and geometric degeneracies coincide (see Appendix A). In that latter case one can then use (27) with straightforward modifications and obtain (28).

The term $\mathcal{O}[e^{-N\nu_1(\mathbf{x}_{N\dots1})}]$ in (27, 28) can be neglected for $N \to \infty$ provided that $\mu_0 > \nu_1(\mathbf{x}_{N\dots1} \in \Omega_N^*)$. The multiplicative correction $\mathcal{O}(1)$ in (28) comes from the eigenvectors in (27). This correction can be neglected if μ_0 stays finite for $N \to \infty$. Below we assume that these two hypotheses hold. This implies from (21) a straightforward relation between the entropy h and the spectral radius $\lambda[\mathbb{T}(\mathbf{x}_{N\dots1})]$ of $\mathbb{T}(\mathbf{x}_{N\dots1})$:

$$h = \mu_0 = \mathrm{lim}_{N\to\infty}\{\, -\frac{1}{N} \ln \lambda[\mathbb{T}(\mathbf{x}_{N\dots1})]\, \}. \tag{29}$$

The relation between the top Lyapunov exponent and the entropy is known [11, 12]. The above discussion emphasizes the role of the spectral radius in this relation [27].

V. GENERATING FUNCTION AND ATYPICAL REALIZATIONS

While the entropy characterizes typical realizations of the process, it is of interest (mainly for a finite number of realizations) to describe atypical realizations, those which fall out of the typical set Ω_N^*.

To this end let us introduce the generating function [8]

$$\Lambda^N(n, N) = \sum_{\mathbf{x}_{N\dots1}} \lambda^n\left[\mathbb{T}(\mathbf{x}_{N\dots1})\right], \tag{30}$$

where n is a non-negative number. (Note that $\Lambda^N(n,N)$ means $\Lambda(n,N)$ in degree of N.)

The generating function $\Lambda^N(n,N)$ is an analog of the partition sum in statistical physics [8] [5]. Writing

$$\Lambda^N(n,N) = \sum_{\mathbf{x}_{N\dots1}\in\Omega_N^*} \lambda^n\left[\mathbb{T}(\mathbf{x}_{N\dots1})\right] + \sum_{\mathbf{x}_{N\dots1}\notin\Omega_N^*} \lambda^n\left[\mathbb{T}(\mathbf{x}_{N\dots1})\right], \tag{31}$$

one notes that in the limits $N\to\infty$ and $n\to1$ the second contribution in the RHS of (31) can be neglected due to definition (21, 23) of the typicality, and then $\Lambda^N(n,N) = \Lambda^N(n) = e^{-(n-1)Nh}$; see (27, 28). Here we already noted that $\Lambda(n,N)$ does not depend on N for $N\to\infty$, and denoted (in this limit) $\Lambda(n,N)=\Lambda(n)$.

Taking into account that $\Lambda(1)=1$, the entropy h is calculated via derivative of the generating function:

$$\begin{aligned} h &= -\frac{1}{N}\,\frac{\partial\Lambda^N(n)}{\partial n}\bigg|_{n=1} = -\,\frac{\mathrm{d}\Lambda(n)}{\mathrm{d}n}\bigg|_{n=1} \equiv -\Lambda'(1) && (32)\\ &= -\sum_{\mathbf{x}_{N\dots1}} \lambda\left[\mathbb{T}(\mathbf{x}_{N\dots1})\right]\ln\lambda\left[\mathbb{T}(\mathbf{x}_{N\dots1})\right]. && (33)\end{aligned}$$

The generating function (30) can be employed for estimating the weight of atypical sequences. This estimate is known as the Chernoff bound [6, 8], and now we briefly recall its derivation adopted to our situation.

Consider the overall weight of atypical sequences, which have probability lower than the typical-sequence probability e^{-Nh}; see (21, 23). These atypical sequences are defined to satisfy

$$-\ln\lambda\left[\mathbb{T}(\mathbf{x}_{N\dots1})\right] > (1+\eta)Nh, \tag{34}$$

where $\eta>0$ quantifies the deviation from the typical behavior. Let $\overline{\sum}_{\mathbf{x}_{N\dots1}}$ be the sum over all those $\mathbf{x}_{N\dots1}$ that satisfy to (34). Define an auxiliary probability distribution $\widetilde{P}(\mathbf{x}_{N\dots1}|n) = \Lambda^{-N}(n,N)\,\lambda^n\left[\mathbb{T}(\mathbf{x}_{N\dots1})\right]$. The sought weight of the atypical sequences is expressed as ($\eta>0$ and $0<n<1$):

$$\begin{aligned}\overline{\sum}_{\mathbf{x}_{N\dots1}}\lambda\left[\mathbb{T}(\mathbf{x}_{N\dots1})\right] &= \Lambda^N(n,N)\overline{\sum}_{\mathbf{x}_{N\dots1}}\widetilde{P}(\mathbf{x}_{N\dots1}|n)\,e^{(1-n)\ln\lambda[\mathbb{T}(\mathbf{x}_{N\dots1})]}\\ &\le e^{N[\ln\Lambda(n,N)+(n-1)(1+\eta)h]}\,\overline{\sum}_{\mathbf{x}_{N\dots1}}\widetilde{P}(\mathbf{x}_{N\dots1}|n) \le e^{N[\ln\Lambda(n,N)+(n-1)(1+\eta)h]}. && (35)\end{aligned}$$

Eq. (35) leads to the following upper (Chernoff) bound for the weight of atypical sequences with the probability lower than the e^{-Nh}:

$$\sum_{-\ln\lambda[\mathbb{T}(\mathbf{x}_{N\dots1})]>(1+\eta)Nh}\lambda\left[\mathbb{T}(\mathbf{x}_{N\dots1})\right] \le e^{-N f(\eta)}, \tag{36}$$

$$f(\eta)\equiv \max{}_{0<n<1}\left[\ln\frac{1}{\Lambda(n)} + (1-n)(1+\eta)h\right], \quad \eta>0. \tag{37}$$

Analogously to (35) we get for the weight of the atypical sequences with the probability higher than the e^{-Nh} ($0<\eta<1$):

$$\sum_{-\ln\lambda[\mathbb{T}(\mathbf{x}_{N\dots1})]<(1-\eta)Nh}\lambda\left[\mathbb{T}(\mathbf{x}_{N\dots1})\right] \le e^{-N g(\eta)}, \tag{38}$$

$$g(\eta)\equiv \max{}_{n>1}\left[\ln\frac{1}{\Lambda(n)} + (1-n)(1-\eta)h\right], \quad \eta>0. \tag{39}$$

The functions $f(\eta)$ and $g(\eta)$ in (37) and (39), respectively, are called the rate functions [6]. It is seen that $f(\eta)$ and $g(\eta)$ are the Legendre transforms of $\ln\Lambda(n)$. The latter is a convex function of n, $\frac{\mathrm{d}^2}{\mathrm{d}^2 n}\ln\Lambda(n)\ge0$, as follow from its definition (30). Then $f(\eta)$ and $g(\eta)$ are convex as well [8]. For example taking into account that n and η are related via the extremum condition $\frac{\mathrm{d}}{\mathrm{d}n}\ln\Lambda(n) = -(1+\eta)h$, we get $f''(\eta) = \left(\frac{\mathrm{d}n}{\mathrm{d}\eta}\right)^2\left[\frac{\mathrm{d}^2}{\mathrm{d}n^2}\ln\Lambda(n)\right]_{n=n(\eta)}\ge0$.

While the above reasoning is based on the Chernoff bounds, there is another (related, but more formal) approach to describing atypical realization, which is known as the measure concentration theory. For a recent application of this theory to HMP see [24].

[5] $\Lambda(n,N)$ is sometimes called the generalized Lyapunov exponent. It is closely related to the concept of multi-fractality [21].

VI. ZETA FUNCTION AND ITS EXPANSION OVER THE PERIODIC ORBITS (CYCLES).

A. Zeta function and entropy.

In this section we show how to adopt the method proposed in [25, 27] for calculating the moment-generating function $\Lambda(n)$ (and thus for calculating the entropy h via (32)). The method is based on the concepts of the zeta-function and periodic orbits.

Define the inverse zeta-function as [8, 25, 26, 28]

$$\xi(z,n) = \exp\left[-\sum_{m=1}^{\infty}\frac{z^m}{m}\Lambda^m(n,m)\right], \tag{40}$$

where $\Lambda^m(n,m) \geq 0$ is given by (30). The analogs of (40) are well-known in the theory of dynamic systems; see [26] for a mathematical introduction, and [25, 27, 28] for a physicist-oriented discussion.

Since for a large N, $\Lambda^N(n,N) \to \Lambda^N(n)$, the zeta-function $\xi(z,n)$ has a zero at $z = \frac{1}{\Lambda(n)}$:

$$\xi(\frac{1}{\Lambda(n)}, n) = 0. \tag{41}$$

Indeed for z close (but smaller than) $\frac{1}{\Lambda(n)}$, the series $\sum_{m=1}^{\infty}\frac{z^m}{m}\Lambda^m(n,m) \to \sum_{m=1}^{\infty}\frac{[z\Lambda(n)]^m}{m}$ almost diverges and one has $\xi(z) \to 1 - z\Lambda(n)$.

Recalling that $\Lambda(1) = 1$ and taking $n \to 1$ in

$$0 = \frac{\mathrm{d}}{\mathrm{d}n}\xi(\frac{1}{\Lambda(n)}, n) = -\frac{\Lambda'(n)}{\Lambda^2(n)}\,\frac{\partial}{\partial z}\xi(\frac{1}{\Lambda(n)}, n) + \frac{\partial}{\partial n}\xi(\frac{1}{\Lambda(n)}, n), \tag{42}$$

we get for the entropy from (32)

$$h = -\Lambda'(1) = -\frac{\frac{\partial}{\partial n}\xi(1,1)}{\frac{\partial}{\partial z}\xi(1,1)}. \tag{43}$$

B. Expansion over the periodic orbits.

In Appendix E 2 we describe following to [25, 26, 27, 28] that under conditions (13, 14) one can expand $\xi(z,n)$ over the periodic orbits:

$$\xi(z,n) = \prod_{p=1}^{\infty}\prod_{\Gamma_p\in\mathrm{Per}(p)}\left[1 - z^p\lambda^n[T(x_{\gamma_1})...T(x_{\gamma_p})]\,\right], \tag{44}$$

$$\Gamma_p \equiv (\gamma_1, ..., \gamma_p), \tag{45}$$

where $\gamma_i = 1, ..., M$ are the indices referring to the realizations of the random process $\mathcal{X}$. The set of periodic orbits $\mathrm{Per}(p)$ contains sequences $\Gamma_p = (\gamma_1, ..., \gamma_p)$ selected according to the following two rules: *i)* Γ_p turns to itself after p successive cyclic permutations of its elements, but it does not turn to itself after any smaller (than p) number of successive cyclic permutations; *ii)* if Γ_p is in $\mathrm{Per}(p)$, then $\mathrm{Per}(p)$ contains none of those $p-1$ sequences obtained from Γ_p under $p-1$ successive cyclic permutations. Concrete examples of $\mathrm{Per}(p)$ for $M = 2, 3$ are given in Tables IV and V.

It is more convenient to present (44) as an infinite sum [25, 27, 29]

$$\xi(z,n) = 1 - z\sum_{l=1}^{M} l_l + \sum_{k=2}^{\infty}\varphi_k(n)z^k, \tag{46}$$

where we defined

$$l_{\alpha...\beta} \equiv l[T(x_\alpha)...T(x_\beta)], \qquad l_{\alpha+\beta} \equiv l[T(x_\alpha)]l[T(x_\beta)], \tag{47}$$

and where $\varphi_k(n)$ are calculated from (44, 45) and recipes presented in Appendix E. These calculations become tedious for large values of k in $\varphi_k(n)$. This is why in Appendix E 3 it is shown how to generate $\varphi_k(n)$ via Mathematica 5.

For two ($M = 2$) realizations of the HMP we employ the notations (47) and get for the first few terms of the product (44) [consult Table IV for understanding the origin of these terms]

$$\xi(z,n) = (1 - z\lambda_1^n)\,(1 - z\lambda_2^n)\,(1 - z\lambda_{12}^n)\,(1 - z\lambda_{122}^n)\,(1 - z\lambda_{112}^n) \tag{48}$$

$$(1 - z\lambda_{1222}^n)\,(1 - z\lambda_{1112}^n)\,(1 - z\lambda_{1122}^n) \prod_{p=5}^{\infty} \prod_{\Gamma_p \in \mathrm{Per}(p)} \left(1 - z^p \lambda_{\gamma_1 \dots \gamma_p}^n\right). \tag{49}$$

For the first six terms of the expansion (46) we get

$$\varphi_2(n) = -l_{12} + l_{1+2}, \tag{50}$$

$$\varphi_3(n) = -l_{221} + l_{2+21} - l_{112} + l_{1+12}, \tag{51}$$

$$\varphi_4(n) = -l_{1122} + l_{2+211} - l_{1222} + l_{2+122} - l_{1112} + l_{1+211} \tag{52}$$

$$-l_{1+2+12} + l_{1+122} \tag{53}$$

$$\varphi_5(n) = -l_{11222} + l_{1+1222} - l_{11122} + l_{2+1112} \tag{54}$$

$$-l_{11112} + l_{1+1112} - l_{12222} + l_{2+1222} \tag{55}$$

$$-l_{12121} + l_{1+1122} - l_{12122} + l_{2+1122} \tag{56}$$

$$-l_{1+2+122} + l_{12+122} - l_{1+2+112} + l_{12+112}, \tag{57}$$

$$\varphi_6(n) = -l_{111122} + l_{1+11122} - l_{112122} + l_{1+12122} - l_{111222} + l_{1+11222} \tag{58}$$

$$-l_{111212} + l_{1+11212} - l_{112222} + l_{1+12222} - l_{222121} + l_{2+22121} \tag{59}$$

$$-l_{122222} + l_{2+12222} - l_{111112} + l_{1+11112} - l_{112212} + l_{2+12121} \tag{60}$$

$$-l_{1+12+122} + l_{1+12122} - l_{2+12+211} + l_{12+1122} - l_{1+12+211} + l_{12+2111} \tag{61}$$

$$-l_{2+12+122} + l_{12+1222} - l_{1+2+1222} + l_{2+11222} - l_{1+2+2111} + l_{2+21111} \tag{62}$$

$$-l_{1+2+1122} + l_{122+211}. \tag{63}$$

In section VII E we study examples, where the expansion (46) can be summed exactly. In these examples the sum in (46) exponentially convergences for $|z| < \alpha^n$, where $\alpha > 1$ is a parameter. As discussed in [28], the exponential convergence of $\xi(z)$ is expected to be a general feature, and it is supported by rigorous results on the structure of the zeta-function.

1. *The structure of $\varphi_k(n)$.*

Note that φ_k consists of even number of terms. The terms are grouped in pairs, e.g., $[-l_{221} + l_{2+21}] + [-l_{112} + l_{1+12}]$ for φ_3, and analogously for other φ_k's. Each pair has the form $-l_A + l_B$, where A and B have the same number of symbols 1 and the same number of symbols 2. This feature ensures that when the spectral radius of the product is equal to the product of the spectral radii, all the terms φ_k will vanish. Ultimately, this is the feature that enforces the convergence of (46) [25, 28]. Once it converges, we can approximate $\xi(z,n)$ by a polynomial of a finite order.

The set of pairs for each φ_k can be divided further into several groups. The first group is formed by (50) and (51) for φ_2 and φ_3, respectively, by (52) for φ_4, by (54–56) for φ_5, and by (58–60) for φ_6. The pairs in this group have the form $-l_{Al} + l_{A+l}$, where $l = 1$ or $l = 2$. If A contains m indices and if m is large, we expect $\ln l_A = \mathcal{O}(m)$ according to the discussion in section IV B. Then

$$-l_{Al} + l_{A+l} \to 0 \qquad \text{for} \qquad m \to \infty. \tag{64}$$

The second group is given by (53) for φ_4, (57) for φ_5, and by (61, 62) for φ_6. In this second group the terms have the form $-l_{A+B+C} + l_{A+BC} = l_A(l_{B+C} - l_{BC})$. Here the term $(l_{B+C} - l_{BC})$ has the structure of the first group. For B or/and C containing a large number of indices, $(l_{B+C} - l_{BC})$ will go to zero.

Finally the third group appears only for $k \geq 6$. For $k = 6$ this group has only one pair given by (63). The members of this third group are of the form $-l_{A+B+CD} + l_{ABD+C}$.

Let us return to (64), which holds, in particular, for A consisting of the same type of indices (e.g., A containing only 1's). Recalling our discussions after (28) and after (63), and expanding A over its eigenvalues and eigenvectors, we conclude *heuristically* that for the convergence radius of $\sum_{k=2}^{\infty}\varphi_k(n)z^k$ in (46) to be sufficiently larger than 1, it is necessary to have for the transfer-matrices $T(x)$ (using notations (12))

$$\lambda[T(x)] \not\approx l_1[T(x)], \quad \lambda[T(x)] \not\approx 1, \tag{65}$$

i.e., closer is $\lambda[T(x)]$ to $l_1[T(x)]$ and or $\lambda[T(x)]$ to 1, more terms are needed in the expansion (46) for the reliable estimate of the entropy. Note that if $\lambda[T(x)] = l_1[T(x)] > l_2[T(x)]$, the first relation in (65) should be modified to $\lambda[T(x)] \not\approx l_2[T(x)]$. We shall meet such examples below; see (81) and the discussion before it.

Recall from (43) that for calculating the entropy we need to know $\xi(z,n)$ in the vicinity of $z=1$ and $n=1$. If the qualitative conditions (65) are satisfied, we expect that the vicinity of $z=1$ and $n=1$ is included in the convergence area. The convergence of expansions similar to (46) is discussed in [25, 27, 28]. In particular, Refs. [25, 27] employ criteria similar to (65) and test them numerically.

In the context of expansion (46) we should mention the results devoted to analyticity properties of the top Lyapunov exponent [30, 31] and of the entropy for HMP [32]. In particular, Ref. [32] states that the entropy h of HMP is an analytic function of the Markov transition probabilities (3), provided that these probabilities are positive. At the moment it is unclear for the present author how *in general* this analyticity result can be linked to the expansion (46). However, we show below on concrete examples that the expansion (46) can be recast into an expansion over the Markov transition probabilities (3).

VII. THE SIMPLEST AGGREGATED MARKOV PROCESS.

A. Definition.

An Aggregated Markov Process (sometimes called a Markov source) is a particular case of HMP, where the probabilities $\pi(x|s)$ in (5) take only two values 0 and 1 [2, 5]. Thus it is defined by the underlying Markov process $\mathcal{S}$ together with a deterministic function $F(s_i)$ that takes the realizations of the Markov process to those of the aggregated process: $\mathcal{X} = (\mathcal{X}_1, \mathcal{X}_2, ...) = (F(\mathcal{S}_1), F(\mathcal{S}_2), ...)$. The function F is not one-to-one so that at least two realizations of $\mathcal{S}$ are lumped together into one realization of $\mathcal{X}$.

The simplest example is given by a Markov process $\mathcal{S} = \{\mathcal{S}_0, \mathcal{S}_1, ...\}$ with three realizations $\mathcal{S}_i = 1,2,3$, such that, e.g., the realizations 2 and 3 of $\mathcal{S}_i$ are not distinguished from each other and correspond to one realization 2 of the observed process $\mathcal{X}_i$ [see Fig. 1]:

$$F(1) = 1, \quad F(2) = F(3) = 2, \tag{66}$$

$$\pi(1|1) = 1, \quad \pi(1|2) = 0, \quad \pi(1|3) = 0, \tag{67}$$

$$\pi(2|1) = 0, \quad \pi(2|2) = 1, \quad \pi(2|3) = 1. \tag{68}$$

The transition matrix of a general three-realization Markov process is [see Fig. 1]

$$\mathbb{P} = \begin{pmatrix} 1-p_1-p_2 & q_1 & r_1 \\ p_1 & 1-q_1-q_2 & r_2 \\ p_2 & q_2 & 1-r_1-r_2 \end{pmatrix}, \qquad |\mathrm{st}\rangle \propto \begin{pmatrix} q_1(r_1+r_2)+q_2r_1 \\ r_2(p_1+p_2)+p_1r_1 \\ p_2(q_1+q_2)+p_1q_2 \end{pmatrix} \tag{69}$$

where all elements of $\mathbb{P}$ are positive, and where we presented the stationary vector $|\mathrm{st}\rangle$ up to the overall normalization [6].

The process $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, ...\}$ has two realizations: $\mathcal{X}_i = 1,2$. The corresponding transfer matrices read from (7)

$$T(1) = \begin{pmatrix} 1-p_1-p_2 & q_1 & r_1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad T(2) = \begin{pmatrix} 0 & 0 & 0 \\ p_1 & 1-q_1-q_2 & r_2 \\ p_2 & q_2 & 1-r_1-r_2 \end{pmatrix}. \tag{70}$$

Note that the second (sub-dominant) eigenvalue of the transfer-matrix product $\mathbb{T}(\mathbf{x}_{N...1}) = \prod_{k=1}^{N} T(x_k)$ (with separate transfer-matrices defined by (70)) is equal to zero, since this eigenvalue can be presented as that of the matrix $T(1)A$,

[6] Note that some authors present the Markov transition matrices $\mathbb{P}$ is such a way that the elements in each raw sum to one. This amounts to transposition of (69). The representation (69) is perhaps more familiar to physicists.

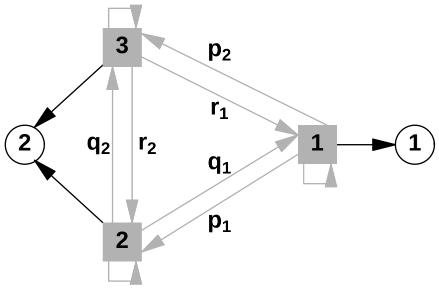

FIG. 1: Schematic representation of the hidden Markov process defined by (66–70). The gray squares and gray arrows indicate, respectively, on the realization of the internal Markov process and transitions between the realizations; see (69). The circles and black arrows indicate on the realizations of the observed process. The gray arrows are probabilistic; the corresponding probabilities are indicated next to them. The black arrow are deterministic; see (66).

where A is some 3×3 matrix. The only exclusion, which has a non-zero sub-dominant eigenvalue, is the realization of $\mathcal{X}$ that does not contain 1 at all: $\mathbb{T}(2...2) = T^N(2)$.

The considered HMP (66–70) belongs to the class of HMP with unambiguous symbol, since the Markov realization 1 is not corrupted by the noise; see Fig. 1. For such HMP, Ref. [32] reports several results on the analytic features of the entropy.

B. Unifilar process.

Before studying in detail the HMP defined by (66–70), let us mention one example of HMP, where the entropy can be calculated directly [2, 5]. This unifilar process is defined as follows [5]: for each realization s_i of the Markov process $\mathcal{S}$ consider realizations s_j with a strictly positive transition probability $p(s_j|s_i) > 0$. Now require that the realizations $F(s_j)$ of $\mathcal{X}_j$ are distinct. Thus given the realization s_i of $\mathcal{S}_1$, there is one to one correspondence between the realizations of $(\mathcal{X}_1, \mathcal{X}_2, ...)$ and those of $(\mathcal{S}_1, \mathcal{S}_2, ...)$. Write the block-entropy of $\mathcal{X}$ as

$$H(\mathcal{X}_N, ..., \mathcal{X}_1) = H(\mathcal{X}_N, ..., \mathcal{X}_1|\mathcal{S}_1) + H(\mathcal{S}_1) - H(\mathcal{S}_1|\mathcal{X}_1, ..., \mathcal{X}_N), \tag{71}$$

where $H(\mathcal{A}|\mathcal{B}) \equiv -\sum_{a,b} \Pr(a,b) \ln \Pr(a|b)$ is the conditional entropy of the stochastic variable $\mathcal{A}$ given $\mathcal{B}$. Due to the definition of the unifilar process: $H(\mathcal{X}_N, ..., \mathcal{X}_1|\mathcal{S}_1) = H(\mathcal{S}_N, ..., \mathcal{S}_2|\mathcal{S}_1)$. The latter is worked out via the Markov feature:

$$H(\mathcal{S}_N, ..., \mathcal{S}_2|\mathcal{S}_1) = (N-1)h_{\rm markov}, \tag{72}$$

$$h_{\rm markov} = -\sum_{k,l} p_{\rm st}(k) p(l|k) \ln p(l|k), \tag{73}$$

where $p_{\rm st}(k)$ is the stationary Markov probability defined in (4), and where $p(l|k)$ are the Markov transition probabilities from (3). Since $H(\mathcal{S}_1)$ and $H(\mathcal{S}_1|\mathcal{X}_1, ..., \mathcal{X}_N)$ in (71) are finite in the limit $N \to \infty$, the entropy $h(\mathcal{X})$ of the unifilar process reduces to that of the underlying Markov process $h_{\rm markov}$ [5].

Note that any finite-order Markov process (conventionally assuming that the usual Markov process is of first order) can be presented as a unifilar process. There are, however, unifilar processes that do not reduce to any finite-order Markov process [5] [7]. The main problem in identifying unifilar processes is that even if $\mathcal{X}$ is not unifilar for given $\mathcal{S}$, it can be still unifilar with respect to another Markov process $\mathcal{S}'$ (see section VII C below for the simplest example). This makes especially difficult the recognition of unifilar processes that do not reduce to any finite-order Markov process.

[7] The example of such a process given in [5] is not minimal. The minimal example is given by four-realization Markov process with non-zero transition probabilities $p(4|1)$, $p(3|4)$, $p(2|3)$, $p(1|2)$, $p(1|1)$, $p(2|2)$, $p(3|3)$ and $p(4|4)$ (all other transition probabilities are zero), and two realizations of $\mathcal{X}_i$ such that $F(1) = F(3) = 1$, $F(2) = F(4) = 2$. The unifilar process $\mathcal{X}$ does not reduce to a finite-order Markov process, since, e.g., there are two different mechanisms of producing the sequence 1...1. This means that $P(1|111)$ is not equal to $P(1|11)$, *etc.*

C. Particular cases.

We now return to the HMP (66–70) and discuss some of its particular cases.

1. For $q_2 = r_2$ and $q_1 = r_1$ all the terms φ_k with $k \geq 3$ in the expansion (46) are zero. One can check that for this case the observed process $\mathcal{X}$ is by itself Markov.

2. For $(1-q_1-q_2)(1-r_1-r_2) = q_2 r_2$, one can check that $\phi_k = 0$ for $k \geq 4$. Now the process $\mathcal{X}$ is the second-order Markov: $P(x_k|x_{k-1}, x_{k-2}, x_{k-3}) = P(x_k|x_{k-1}, x_{k-2})$.

Thus at least for these two cases the calculation of the entropy is straightforward.

The above two facts tend to clarify the meaning of the expansion (46). It is tempting to suggest that if the expansion (46) is cut precisely at a positive integer $K > 2$, i.e., $\varphi_{k\geq K} = 0$, then the corresponding process $\mathcal{X}$ is $K-2$-order Markovian. If true, this will give convenient conditions for deciding on the finite-order Markov feature, and will mean that the successive terms in (46) are in fact approximations the HMP via finite-order Markov processes.

D. Upper and lower bounds for the entropy.

Before presenting the main results of this section, let us recall that the entropy of any (stationary) HMP satisfies the following inequalities [6] [8]:

$$H(\mathcal{X}_2|\mathcal{S}_1) \leq h \leq H(\mathcal{X}_2|\mathcal{X}_1) \equiv H(2) - H(1), \tag{74}$$

where $H(\mathcal{A}|\mathcal{B}) = -\sum_{a,b} \Pr(\mathcal{A} = a, \mathcal{B} = b) \ln \Pr(\mathcal{A} = a|\mathcal{B} = b)$ and $H(N)$ are, respectively, the conditional entropy and the block entropy defined in (16). Employing (5, 7) we deduce

$$\Pr(\mathcal{X}_2 = x|\mathcal{S}_1 = s) = \sum_{s'=1}^{L} T_{s'\,s}(x). \tag{75}$$

This equation together with the stationary probability (69) of the Markov process is sufficient for calculating $H(\mathcal{X}_2|\mathcal{S}_1)$ for the HMP (66, 70):

$$H(\mathcal{X}_2|\mathcal{S}_1) = p_{\rm st}(1)\chi(p_1+p_2) + p_{\rm st}(2)\chi(q_1) + p_{\rm st}(3)\chi(r_1), \tag{76}$$

$$\chi(p) \equiv -p\ln p - (1-p)\ln(1-p). \tag{77}$$

The upper bound $H(\mathcal{X}_2|\mathcal{X}_1)$ is calculated directly from (8, 10, 16).

E. Generating function and entropy: exact results.

For a particular four-parametric class of HMP (66–70) we were able to sum exactly the expansion (46) [9]. This class is characterized by the condition that the two leading eigenvalues of the transfer-matrix $T(2)$ in (70) have equal absolute values [the third eigenvalue is equal to zero]:

$$\lambda[T(2)] = \lambda_1[T(2)]. \tag{78}$$

A direct inspection shows that this condition amounts to two possible forms (80) and (88) of the transition matrix $\mathbb{P}$. These two cases are studied below.

1. First case.

For this first case the transition matrix is obtained from (70) under [10]

$$r_2 = 0 \quad \text{and} \quad r_1 = q_1 + q_2. \tag{79}$$

[8] Eq. (74) is a particular case of a slightly more general inequality [6, 10]. For our purely illustrative purposes (74) is sufficient.

[9] This was done by hands, checking the separate terms of the expansion (44).

[10] Or, alternatively, via $q_2 = 0$ and $q_1 = r_1 + r_2$. This, however, does not amount to anything new as compared to (80).

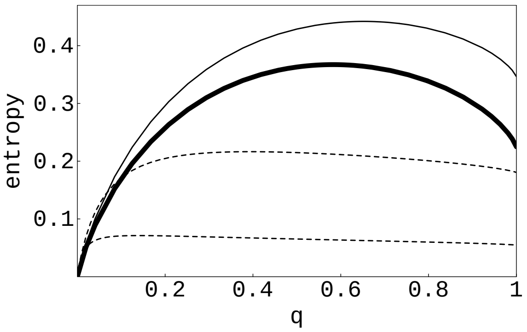

FIG. 2: Entropy (85) of HMP (69, 70, 80) versus $q = q_2$ for $p_2 = q_1 = 0$. Normal line: $p_1 = 0.5$. Thick line: $p_1 = 0.75$. Upper dashed line: $p_1 = 0.05$. Lower dashed line: $p_1 = 0.01$. It is seen that for a small value of p_1, the entropy h is nearly constant for a range of $q = q_2$.

This leads from (69) to the transition matrix

$$\mathbb{P} = \begin{pmatrix} 1 - p_1 - p_2 & q_1 & q_1 + q_2 \\ p_1 & 1 - q_1 - q_2 & 0 \\ p_2 & q_2 & 1 - q_1 - q_2 \end{pmatrix}. \tag{80}$$

It is seen that the realization $\{\mathcal{S}_{k+1} = 2,\, \mathcal{S}_k = 3\}$ for the Markov process is prohibited. For the HMP there are no prohibited sequences.

The inverse zeta-function reads from (46):

$$\begin{aligned} \xi(z, n) \;&=\; 1 - [\,(1 - p_1 - p_2)^n + (1 - q_1 - q_2)^n\,]\, z \\ &+\; [\,(1 - p_1 - p_2)^n (1 - q_1 - q_2)^n - (p_1 q_1 + p_2(q_1 + q_2)\,)^n\,] z^2 \\ &+\; z^3 [p_1 q_2 (q_1 + q_2)]^n\, [\,\Phi(y, -n, b) - \Phi(y, -n, b + 1)\,]\,, \end{aligned} \tag{81}$$

where we defined

$$b \equiv (1 - q_1 - q_2) \frac{p_2(q_1 + q_2) + p_1 q_1}{p_1 q_2 (q_1 + q_2)}, \tag{82}$$

$$y \equiv (1 - q_1 - q_2)^n z, \tag{83}$$

and where $\Phi(y, -n, b)$ is the Lerch Φ-function:

$$\Phi(y, -n, b) = \sum_{k=0}^{\infty} (k + b)^n y^k. \tag{84}$$

In this representation, which led to (81), the sum converges for $|y| < 1$ or for $z < (1 - q_1 - q_2)^{-n} \geq 1$. The convergence radius tends to one for $q_1 + q_2 \to 0$, or, equivalently, for $\lambda[T(2)] \to 1$; see (70). This violates the second qualitative condition in (65).

Using (43) we get from (81) for the entropy:

$$\begin{aligned} &h = -\frac{1}{p_1 + p_2 + q_1 + q_2 + \frac{p_1 q_2}{q_1 + q_2}} \{ \\ &(1 - p_1 - p_2)(q_1 + q_2) \ln(1 - p_1 - p_2) + (1 - q_1 - q_2)(p_1 + p_2 + \frac{p_1 q_2}{q_1 + q_2}) \ln(1 - q_1 - q_2) \\ &+ p_1 q_2 \ln[\,p_1 q_2 (q_1 + q_2)\,] + [\,(p_1 + p_2) q_1 + p_2 q_2\,] \ln[\,(p_1 + p_2) q_1 + p_2 q_2\,] \\ &+ p_1 q_2 (q_1 + q_2) \Big[\, \Phi'_{[2]}(1 - q_1 - q_2, -1, b) - \Phi'_{[2]}(1 - q_1 - q_2, -1, b + 1) \Big] \}, \end{aligned} \tag{85}$$

where b is defined in (82), and where

$$\Phi'_{[2]}(y, -1, b) = \sum_{k=0}^{\infty} \ln \left[\frac{1}{k + b} \right] (k + b) y^k. \tag{86}$$

TABLE I: For two set of parameters of the HMP (66, 79, 80) we present the exact value of entropy h obtained from (85), the lower bound $H(\mathcal{X}_2|\mathcal{S}_1)$, and the upper bound $H(\mathcal{X}_2|\mathcal{X}_1)$; see (74).

	h	$H(\mathcal{X}_2\|\mathcal{S}_1)$	$H(\mathcal{X}_2\|\mathcal{X}_1)$
$p_1 = 0.75$ $p_2 = 0.10$ $q_1 = 0.25$ $q_2 = 0.20$	0.569580	0.557243	0.572373
$p_1 = 0.30$ $p_2 = 0.20$ $q_1 = 0.55$ $q_2 = 0.10$	0.684796	0.682486	0.684843

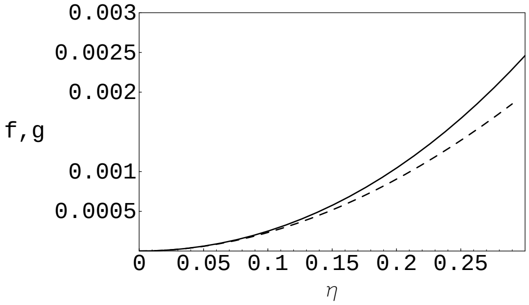

FIG. 3: The rate functions $f(\eta)$ and $g(\eta)$ defined by (37) and (39), respectively for the HMP given by (70, 88, 89). Normal line : $g(\eta)$. Dashed line : $f(\eta)$. For the parameters in (88) we take: $p_1 = 0.2$, $p_2 = 0.3$, $q = 0.05$, and $r = 0.01$. For these values the entropy (90) is $h = 0.166671$.

The behavior of h is illustrated in Fig. 2 for particular values of p_1, p_2, q_1 and q_2. Table I compares the exact expression (85) with the upper and lower bounds (74).

The analytic features of h given by (85) as a function of the Markov transition probabilities p_1, p_2, q_1 and q_2, agree with the results obtained in [32]. In particular, note that for $p_1 + p_2 \to 1$ the entropy h becomes non-analytic due to the term $\propto (1 - p_1 - p_2)\ln(1 - p_1 - p_2)$.

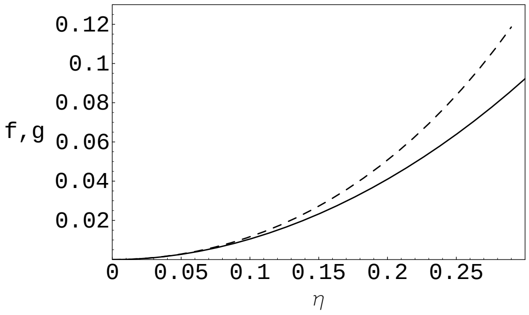

FIG. 4: The same as in Fig. 3 but with $q = 0.1$ and $r = 0.4$. For these values the entropy (90) is $h = 0.619519$, which is larger than the entropy in Fig. 3.

TABLE II: For two set of parameters of the HMP (69, 70, 88) we present the exact value of entropy h obtained from (90), the lower bound $H(\mathcal{X}_2|\mathcal{S}_1)$, and the upper bound $H(\mathcal{X}_2|\mathcal{X}_1)$; see (74). The parameters p_1, p_2, q and r are tuned such that $H(\mathcal{X}_2|\mathcal{S}_1)$ and $H(\mathcal{X}_2|\mathcal{X}_1)$ provide rather tight bounds on h.

	h	$H(\mathcal{X}_2\|\mathcal{S}_1)$	$H(\mathcal{X}_2\|\mathcal{X}_1)$
$p_1 = 0.1$ $p_2 = 0.1$ $q = 0.2$ $r = 0.3$	0.528531	0.525571	0.528534
$p_1 = 0.2$ $p_2 = 0.2$ $q = 0.3$ $r = 0.4$	0.659897	0.656974	0.659901

2. Second case.

The second possibility of satisfying (78) is given by

$$q_1 + q_2 = 1 \quad \text{and} \quad r_1 + r_2 = 1, \tag{87}$$

$$\mathbb{P} = \begin{pmatrix} 1-p_1-p_2 & q & r \\ p_1 & 0 & 1-r \\ p_2 & 1-q & 0 \end{pmatrix}. \tag{88}$$

The realizations of the corresponding Markov process do not contain $\{\mathcal{S}_{k+1} = 2,\, \mathcal{S}_k = 2\}$ and $\{\mathcal{S}_{k+1} = 3,\, \mathcal{S}_k = 3\}$. Again, the realizations of the HMP do not have any prohibited sequence.

The inverse zeta-function reads from (46)

$$\begin{aligned} \xi(z,n) &= 1 - \Big[(1-p_1-p_2)^n + (1-q)^{n/2}(1-r)^{n/2} \Big]\, z + \Big[-(p_1q+p_2r)^n + (1-p_1-p_2)^n(1-q)^{n/2}(1-r)^{n/2} \Big]\, z^2 \\ &+ \frac{z^3}{1+z(1-q)^{n/2}(1-r)^{n/2}} \Big[(p_1q+p_2r)^n(1-q)^{n/2}(1-r)^{n/2} - (\, p_1r(1-q) + p_2q(1-r)\,)^n \Big]\,. \end{aligned} \tag{89}$$

The series that led to (89) converges for $|z| < (1-q)^{-n/2}(1-r)^{-n/2}$. Again the convergence radius going to one violates the second qualitative condition in (65).

Eqs. (32, 89) imply for the source entropy:

$$\begin{aligned} h = -\frac{1}{2(p_1+p_2)+q(1-p_1)+r(1-p_2)-qr} &\{\, [\, q(1-r)+r\,](1-p_1-p_2)\ln(1-p_1-p_2) \\ &+(p_1+p_2)\,(1-q)\,(1-r)\,\ln[(1-q)(1-r)] + (p_1q+p_2r)\ln[p_1q+p_2r] \\ &+[p_2q(1-r)+p_1(1-q)r]\ln[p_2q(1-r)+p_1(1-q)r]\,\}\,. \end{aligned} \tag{90}$$

Applying the general definition (73) of the Markov entropy to the particular case (69) we get for the Markov entropy

$$\begin{aligned} h_{\rm markov} &= -\frac{1}{2(p_1+p_2)+q(1-p_1)+r(1-p_2)-qr}\,\{ \\ &[\, q(1-r)+r\,][\,(1-p_1-p_2)\ln(1-p_1-p_2) + p_1\ln p_1 + p_2\ln p_2\,] \\ &[(1-r)(p_1+p_2)+p_1r]\,[q\ln q + (1-q)\ln(1-q)] \\ &[p_2+p_1(1-q)]\,[r\ln r + (1-r)\ln(1-r)]\,\}. \end{aligned} \tag{91}$$

Comparing (90, 91) one can check [e.g., numerically] that $h_{\rm markov} > h$, as should be, since lumping several states together decreases the entropy. Table II compares the exact value (90) for the entropy with the upper and lower bounds (74).

3. Rate functions for large deviations.

Recall that the rate function $f(\eta)$ ($g(\eta)$) defined in section V, describe the weight of atypical sequences with the probability smaller (larger) than the typical sequence probability e^{-Nh}. The positive parameter η defines the amount of this smallness (largeness); see (37) and (39).

The calculation of $f(\eta)$ and $g(\eta)$ for the considered HMP model (88, 70) is straightforward. One finds out the zero of the ξ-function given by (89). This will define, via (41), the moment-generating function $\Lambda(n)$. If there are several zeros of $\xi(z,n)$ as a function of z, we select the one that goes to $z=1$ for $n\to 1$. Then $f(\eta)$ and $g(\eta)$ are calculated from their definitions (37) and (39).

The behavior of $f(\eta)$ and $g(\eta)$ as functions of η is presented in Figs. 3 and 4. For each figure we take different sets of parameters p_1, p_2, q and r; see (88) for their definition. To make this difference explicit let us denote $f_3(\eta)$, $g_3(\eta)$ and $f_4(\eta)$, $g_4(\eta)$ for Fig. 3 and Fig. 4, respectively.

Now let us observe that

$$f_3(\eta) < f_4(\eta), \qquad g_3(\eta) < g_4(\eta), \tag{92}$$

$$g_3(\eta) > f_3(\eta), \qquad g_4(\eta) < f_4(\eta). \tag{93}$$

For explaining these inequalities we note that for the parameters of Fig. 3 the entropy is smaller than h in Fig. 4:

$$h_3 < h_4, \tag{94}$$

which means that the typical set Ω_N^* for Fig. 4 contains more sequences, so there remains less of them outside, which may explain (92). For the same reason (94), the probability of each typical sequence is higher for the parameters in Fig. 3. Thus for the parameters presented in Fig. 3 more high-probability sequences are included in the corresponding typical set Ω_N^*. This may explain (93).

In further numerical checkings it was noted that the above relation between (92) and (93) from one side, and (94) from another side, seems to be much more general than these particular examples.

VIII. BINARY SYMMETRIC HIDDEN MARKOV PROCESS.

A. Definition and symmetries.

This is another popular (and simple to define) example of HMP. Now the Markov process has two states 1 and 2. The realizations of the observed (Hidden Markov) process also take two values 1 and 2. The internal Markov process is driven by the conditional probability

$$\mathbb{P} = \begin{pmatrix} p(1|1) & p(1|2) \\ p(2|1) & p(2|2) \end{pmatrix} = \begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix}. \tag{95}$$

The stationary probability for this Markov process is found via (4): $p_{\rm st}(1) = p_{\rm st}(2) = \frac{1}{2}$.

The probabilities for the observations 1 or 2 given the internal state read

$$\pi(x_i|s_i) = \begin{pmatrix} \pi(1|1) & \pi(1|2) \\ \pi(2|1) & \pi(2|2) \end{pmatrix} = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}, \tag{96}$$

where ϵ is the error probability during the observation.

For the transfer matrices we have:

$$T(1) = \begin{pmatrix} \epsilon(1-q) & \epsilon q \\ (1-\epsilon)q & (1-\epsilon)(1-q) \end{pmatrix}, \qquad T(2) = \begin{pmatrix} (1-\epsilon)(1-q) & (1-\epsilon)q \\ \epsilon q & \epsilon(1-q) \end{pmatrix}. \tag{97}$$

$T(2)$ is obtained from $T(1)$ via $\epsilon \to 1-\epsilon$.

TABLE III: For two sets of the parameters q and ϵ of the binary symmetric HMP (95, 96, 97) we present the entropy h obtained by approximating (46) via a polynomial or order 2, 13 and 12, respectively. These values are denoted by h_2, h_{13} and h_{12}. We compare h_k with the lower bound $H(\mathcal{X}_2|\mathcal{S}_1)$, and the upper bound $H(\mathcal{X}_2|\mathcal{X}_1)$; see (74). It is seen that the relative difference $\frac{h_{13}-h_2}{h_{13}}$ is not larger than 0.02.

	h_2	h_{13}	h_{12}	$H(\mathcal{X}_2\|\mathcal{S}_1)$	$H(\mathcal{X}_2\|\mathcal{X}_1)$
$q = 0.2$ $\epsilon = 0.45$	0.687811	0.693108	0.693100	0.691346	0.693129
$q = 0.25$ $\epsilon = 0.4$	0.681322	0.692884	0.692881	0.688139	0.692947

The following symmetry features are deduced directly from (95–97):

(1) For any N the probability $P(x_N, \dots, x_1; q, \epsilon)$ of the binary symmetric HMP is invariant with respect to $\epsilon \to 1-\epsilon$: $P(x_N, \dots, x_1; q, \epsilon) = P(x_N, \dots, x_1; q, 1-\epsilon)$.

(2) The probability $P(x_N, \dots, x_1; q, \epsilon)$ is invariant with respect to the full "inversion" of the realization $(x_N, \dots, x_1)$, e.g. $P(1, 2, 1, 1; q, \epsilon) = P(2, 1, 2, 2; q, \epsilon)$.

(3) In general, the probability $P(x_N, \dots, x_1; q, \epsilon)$ is not invariant with respect to $q \to 1-q$, e.g., $P(1, 2; q, \epsilon) - P(1, 2; 1-q, \epsilon) = \frac{1}{2}(1-2\epsilon)(2q-1)$. However, for each given realization $(x_N, \dots, x_1)$ one can find another unique realization $(\bar{x}_N, \dots, \bar{x}_1)$ such that $P(x_N, \dots, x_1; q, \epsilon) = P(\bar{x}_N, \dots, \bar{x}_1; 1-q, \epsilon)$. The logics of relating $(x_N, \dots, x_1)$ to $(\bar{x}_N, \dots, \bar{x}_1)$ should be clear from the following example: if $(x_4, \dots, x_1) = (1, 2, 2, 1)$, then $(\bar{x}_4, \dots, \bar{x}_1) = (2, 2, 1, 1)$. In more detail, $\bar{x}_4 = 2$ is defined to be different from $x_4 = 1$, and once $x_3 = 2$ is different from $x_4 = 1$, $\bar{x}_3 = 2$ does not differ from $\bar{x}_4 = 2$, *etc.* It should be clear (e.g., by induction) that for a given $(x_N, \dots, x_1)$, $(\bar{x}_N, \dots, \bar{x}_1)$ is indeed unique.

This feature means, in particular, that the entropy h of the binary symmetric HMP—being according to (16, 17) a symmetric function of all probabilities $P(x_N, \dots, x_1)$—is invariant with respect to $q \to 1-q$: $h(q, \epsilon) = h(1-q, \epsilon)$, in addition to being invariant with respect to $\epsilon \to 1-\epsilon$.

(4) In general, the probabilities $P(x_N, \dots, x_1)$ are not invariant with respect to a cyclic interchange of the realizations, e.g., $P(1, 2, 1; q, \epsilon) - P(1, 1, 2; q, \epsilon) = \frac{1}{2}(1-2\epsilon)^2 q(2q-1)$.

For the considered binary symmetric HMP we did not find any exactly solvable situation. Thus, we employed (46) and calculated $\xi(z, n)$ by approximating the infinite sum in the RHS of (46) via a polynom of order K: $\sum_{k=2}^{K} \varphi_k(n) z^k$ [11]. This approximation was suggested in [25] and it is based on the fact that the sum supposed to converge exponentially at least in the vicinity of $z = 1$ and $n = 1$. This is what we saw for the exactly solvable situations (81) and (89). The qualitative criterion for the exponential converges was suggested in [25, 27] and was discussed by us around (65). Since both transfer-matrices in (97) have the same eigenvalues

$$\frac{1}{2}\left[1 - q \pm \sqrt{q^2 + (1-2q)(1-2\epsilon)^2}\right], \tag{98}$$

for the studied binary symmetric HMP there are several cases, where the [qualitative] conditions (65) are violated: *i)* $q \to 0$ and $\epsilon \to \frac{1}{2}$; *ii)* $q \to 1$; *iii)* $q \to 0$ and $\epsilon \to 0$. In these three cases we expect that that approximating $\xi(z, n)$ by $\sum_{k=2}^{K} \varphi_k(n) z^k$ will not be feasible, since large values of K will be required to achieve a reasonably high precision. Fig. 5 and Table III present the results for the entropy obtained in the above approximate way and compare them with the upper and lower bounds, as given by (74).

B. Small-noise limit.

For $\epsilon = \frac{1}{2}$ or for $q = \frac{1}{2}$ the process becomes memory-less: $P(x_1, ..., x_N) = P(x_1)...P(x_N)$. Here all the functions φ_k in (46) are equal to zero. Another particular case is the limit $\epsilon \to 0$ (no noise), where the hidden Markov process degenerates into the original Markov process. It is straightforward to check that in (46) for the entropy only the term ϕ_2 is different from zero, while $\phi_k = 0$ for $k \geq 3$. This produces the well-known expression (73) for the entropy of a Markov process.

[11] The terms in this expansion can perhaps be re-arranged so as to facilitate the convergence. Since in the present paper the numerical calclations serve mainly illustrative purposes, we shall not dwell into this aspect.

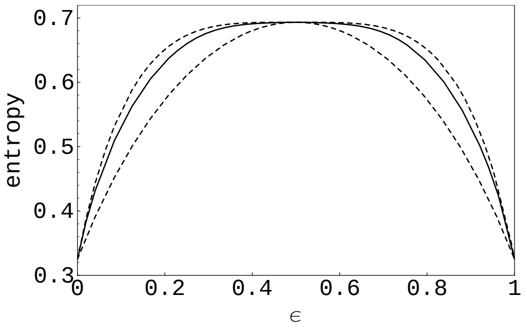

FIG. 5: Entropy of the binary hidden Markov chain (normal line) versus the error probability ϵ for $q = 0.1$. Dashed lines: upper and lower bounds for the entropy as given by (74). The entropy is calculated from (46, 43) approximating the infinite sum in (46) by a poynomial or the order 13.

Let us work out the vicinity of $\epsilon = 0$, assuming that ϵ is small (quasi-Markov situation). One can check that

$$\varphi_k = \mathcal{O}(\epsilon^{k-2}) \qquad \text{for} \qquad k \geq 3. \tag{99}$$

Thus for finding the entropy and the generating function within the order $\mathcal{O}(\epsilon^2)$, we need to expand φ_k with $k = 1, 2, 3, 4$ over ϵ and select all the terms of order $\mathcal{O}(\epsilon)$ and $\mathcal{O}(\epsilon^2)$. We write down explicitly the approximation of $\xi(z, n)$ via the polynom of order 4 (higher-order terms $\varphi_{k\geq 5}$ are not needed, since they do not contribute to the order $\mathcal{O}(\epsilon^2)$):

$$\xi(z, n) = 1 + z\,\varphi_1(n) + z^2\varphi_2(n) + z^3\varphi_3(n) + z^4\varphi_4(n) + \mathcal{O}(z^3). \tag{100}$$

Using (98) and (50–53) we get after straightforward algebraic calculations (taking for simplicity $q < \frac{1}{2}$)

$$\begin{aligned} \varphi_1(n) &= -2\,(1-q)^n \\ &+ \ 2\,\epsilon\, n\,(1-q)^{n-2}\,(1-2\,q) \\ &- \ \epsilon^2\, n\,(1-q)^{n-4}\,(1-2\,q)\,\{\,(1-2q)(n-1-q)+q\,\} + \mathcal{O}(\epsilon^3), \end{aligned} \tag{101}$$

$$\begin{aligned} \varphi_2(n) &= (1-q)^{2\,n} - q^{2\,n} \\ &- \ 2\,\epsilon\, n\,(1-2\,q)\,\left[(1-q)^{2\,(n-1)} + q^{2\,(n-1)}\right] \\ &- \ \epsilon^2\, n\,(1-2\,q)\,\Big[q^{2\,(n-2)}\,\{\,(1-2q)(q+2n-3)-q\,\} \\ &\qquad\qquad + (1-q)^{2(n-2)}\,\{\,(1-2q)(q+1-2n)-q\,\}\Big] + \mathcal{O}(\epsilon^3), \end{aligned} \tag{102}$$

$$\begin{aligned} \varphi_3(n) &= 2\epsilon\, n\,(1-2q)^2\,(1-q)^{n-2}\,q^{2(n-1)} \\ &- \ \epsilon^2\, n\,(1-2q)^2\,(1-q)^{n-4}\,q^{2(n-2)}\,\big[\,5-3n+4q(3n-5)+2q^2(16-7n) \\ &\qquad\qquad + 4q^3(n-6)+10q^4\,\big] + \mathcal{O}(\epsilon^3), \end{aligned} \tag{103}$$

$$\varphi_4(n) = \epsilon^2 n\,(1-2q)^3\,(1-q)^{2(n-2)}\,q^{2(n-2)}\,\big[\,2-4q(1-q)-n(1-2q)\,\big] + \mathcal{O}(\epsilon^3). \tag{104}$$

Note that all ϵ corrections nullify for $q = \frac{1}{2}$, once in this limit we should get a memory-less process. These equations produce for the entropy from (100, 43):

$$h = -(1-q)\ln(1-q) - q\ln q \tag{105}$$

$$- \ 2\epsilon\,(1-2q)\,\ln\left(\frac{1-q}{q}\right) \tag{106}$$

$$- \ 2\epsilon^2\,(1-2q)\left[\ln\left(\frac{1-q}{q}\right) + \frac{1-2q}{4(1-q)^2\,q^2}\right] + \mathcal{O}(\epsilon^3). \tag{107}$$

Eq. (105) is just the Markov entropy (73) obtained in the limit $\epsilon = 0$. Eqs. (106) is the first correction to the Markov situation; it is obtained in [11, 13]. The second correction (107) is reported in [15]. The authors of [15] also obtain the higher-order corrections employing the mapping of the binary symmetric HMP to the one-dimensional Ising model. These higher-order correction can be also obtained within the present method. Thus we demonstrated that the small-noise (quasi-Markov) situation can be adequately explored with the present method.

In addition we obtain the small-noise expressions (101-104) for the zeta-function. This result is new and it allows to find the moment-generating function, which contains more information than the entropy, e.g., (100–104) can be used for approximating the rate functions (37) and (39). In particular, for the generating function we get from (41) and (101–104)

$$\Lambda(n) = q^n + (1-q)^n - \frac{\epsilon n(1-2q)\left[\,(1-q)^{2n}q^2 - (1-q)^2 q^{2n}\,\right]}{q^2(1-q)^2\left[\,(1-q)^n + q^n\,\right]} + \mathcal{O}(\epsilon^2). \tag{108}$$

IX. SUMMARY.

In this paper we studied the entropy and the moment-generating function of Hidden Markov Processes (HMP). The fact that these processes model non-Markov memory is at the origin of their numerous applications, and, simultaneously, the main reason of difficulties in characterizing their entropy and the moment-generating function. Recall that the entropy gives the number of sequences in the typical set of the random process [6, 8]; the typical set is the smallest set of realizations with the overall probability close to one. Alternatively, the entropy is the uncertainty [per time-unit] of the process given its long history. The generating function allows to estimate the [small] probability of atypical sequences via the Chernoff bound and the rate functions [6, 8]. The entropy of HMP was studied via upper and lower bounds [6, 10], expansions over small parameters [15, 16, 17], and via expressing the entropy as a solution of an integral equation [7, 8, 10, 11, 12, 13, 14].

Here we proposed to calculate the entropy and the moment-generating function of HMP via the cycle expansion of the zeta-function, a method adopted from the theory of dynamical systems [25, 27, 29]. I show that this method has two basic advantages. First, it produces exact results, both for the entropy and the moment-generating function, for a class of HMP. We did not so far got into any systematic way of searching for the exact solutions within this method. The examples of exact solutions presented in section VII E were obtained in the most straightforward way. Second, even if no exact solution is found, the method offers an expansion for the entropy and the moment-generating function via an exponentially convergent power series [25, 27, 29]. Cutting off these expansions at some finite order gives normally an improvable approximation for the sought quantities, especially since there are qualitative estimates for the convergence radius of the series. This was demonstrated in section VIII.

As a by-product of this study, we conjectured in section VII C on tentative conditions under which HMP reduces to a finite-order Markov process. These conditions compare favorably with those existing in literature, see e.g. [34], and they deserve further exploration. We also conjectured relations (92–94) between the rate functions of the random process and its entropy.

Acknowledgments

I thank David Saakian for arousing my interest in this problem.

The work was supported by Volkswagenstiftung grant ”Quantum Thermodynamics: Energy and information flow at nanoscale”.

[1] L. R. Rabiner, Proc. IEEE, **77**, 257-286, (1989).
[2] Y. Ephraim and N. Merhav, IEEE Trans. Inf. Th., **48**, 1518-1569, (2002).
[3] M. Crouse, R. Nowak and R. Baraniuk, IEEE Tran. Signal Process., **46**, 886 (1998).
[4] T. Koski, *Hidden Markov Models for Bioinformatics* (Kluwer, Academic Publishers, Dordrecht, 2001).
P. Baldi and S. Brunak, *Bioinformatics* (MIT Press, Cambridge, USA, 2001).
[5] R. Ash, *Information Theory* (Interscience Publishers, NY, 1965).
[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, (Wiley, New York, 1991).
[7] D. Blackwell, *The entropy of functions of finite-state markov chains*, in Trans. First Prague Conf. Inf. Th., Statistical Decision Functions, Random Processes, p. 13 (Pub. House Chechoslovak Acad. Sci., Prague, Czechoslovakia, 1957).
[8] R.L. Stratonovich, *Information Theory* (Sovietskoe Radio, Moscow, 1976) (In Russian).

[9] M. Rezaeian, *Hidden Markov Process: A New Representation, Entropy Rate and Estimation Entropy*, arXiv:cs.IT/0606114.

[10] I.J. Birch, Ann. Math. Stat. **33**, 930 (1962).

[11] P. Jacquet, G. Seroussi, and W. Szpankowski, *On the Entropy of a Hidden Markov Process*, Int. Symp. Inf. Th. p. 10, Chicago, IL, 2004.

[12] T. Holliday, A. Goldsmith and P. Glynn, IEEE Trans. Inf. Th. **52**, 3509 (2006).

[13] E. Ordentlich and T. Weissman, IEEE Trans. Inf. Th., **52**, 19 (2006).

[14] S. Egner *et al.*, *On the Entropy Rate of a Hidden Markov Model*, Int. Symp. Inf. Th., p. 12, Chicago, IL (2004).

[15] O. Zuk, I. Kanter and E. Domany, J. Stat. Phys. **121**, 343 (2005).

[16] O. Zuk, I. Kanter, E. Domany and M. Aizenman, IEEE Signal Processing Letters, **13**, 517 (2006).

[17] P.Chigansky, *The Entropy Rate of a Binary Channel with Slolwly Varying Input*, arXiv:cs/0602074.

[18] R. A. Horn and C. R. Johnson, *Matrix Analysis* (Cambridge University Press, New Jersey, USA, 1985).

[19] J.F.C. Kingman, Ann. Probab. **1**, 883 (1973).

[20] J.M. Steele, Annales de l'I.H.P. B, **25**, 93 (1989).

[21] A. Crisanti, G. Paladin and A. Vulpiani, *Products of Random Matrices in Statistical Physics*, Springer Series in Solid State Sciences, Vol. 104, (Springer, Berlin, 1993).

[22] L.Y. Goldsheid and G.A. Margulis, Russ. Math. Surveys **44**, 11 (1989).

[23] S.A. Orszag, P.L. Sulem and I. Goldirsch, Physica D **27**, 311 (1987).

[24] L. Kontorovich, *Measure Concentration of Hidden Markov Processes*, arXiv:math/0608064.

[25] R. Artuso. E. Aurell and P. Cvitanovic, Nonlinearity **3**, 325 (1990).
P. Cvitanovic, Phys. Rev. Lett. **61**, 2729 (1988).

[26] D. Ruelle, *Statistical Mechanics, Thermodynamic Formalism*, (Reading, MA: Addison-Wesley, 1978).

[27] R. Mainieri, Chaos **2**, 91 (1992).

[28] E. Aurell, J. Stat. Phys., **58**, 967 (1990).

[29] J. Nielsen, *Lyapunov exponents for products of random matrices*, preprint available at http://citeseer.ist.psu.edu/438423.html.

[30] L. Arnold, V. M. Gundlach and L. Demetrius, Ann. Appl. Probab., **4**, 859 (1994).

[31] Y. Peres, Ann. Inst. H. Poincare Probab. Statist., **28**, 131 (1992).

[32] G. Han and B. Markus, IEEE Trans. Inf. Th., **52**, 5251, (2006).

[33] I learned about the function `ListNecklaces2` from the e-mail exchange presented in http://forums.wolfram.com/student-support/topics/6401

[34] L. Gurvits and J Ledoux, Linear Algebra and Applications, **404**, 85 (2005).

[35] K. Petersen, *Lectures on Ergodic Theory*, available from http://www.math.unc.edu/Faculty/petersen/lecturespdf.pdf

APPENDIX A: RECOLLECTION OF SOME FACTS ABOUT THE EIGEN-REPRESENTATION VERSUS SINGULAR VALUE DECOMPOSITION.

A matrix A can be diagonalized if [18]

$$A = V\,D\,V^{-1}, \tag{A1}$$

where D is a diagonal matrix, and where V is an arbitrary invertible matrix. Writing the eigen-resolution of D, $D = \sum_k \alpha_k |\alpha_k\rangle\langle\alpha_k|$, where $\langle\alpha_k|\alpha_n\rangle = \delta_{kn}$, one gets

$$A = \sum_k \alpha_k |R_k\rangle\langle L_k|, \tag{A2}$$

where α_k are the eigenvalues of A (i.e., the solutions of $\det(A - \alpha\,1) = 0$), and where $|R_k\rangle$ and $|L_k\rangle$ are, respectively, the right and left eigenvectors:

$$A|R_k\rangle = \alpha_k|R_k\rangle, \qquad \langle L_k|A = \alpha_k\langle L_k|, \qquad \langle L_k|R_n\rangle = \delta_{kn}. \tag{A3}$$

Note that in general $\langle L_k|L_n\rangle \neq \delta_{kn}$. The right and left eigenvectors coincide for normal matrices $[A, A^\dagger] = 0$ (A commutes with its complex conjugate). For those matrices V is unitary.

Not every matrix can be diagonalized, a necessary and sufficient condition for this is that for each eigenvalue the algebraic degeneracy (i.e., degeneracy of this eigenvalue as the root of the characteristic polynom) coincides with the geometric degeneracy (the number of eigenvectors corresponding to this eigenvalue; geometric degeneracy cannot be larger than the algebraic one). Thus a sufficient condition for a matrix to be diagonalizable is that its eigenvalues are not degenerate. Here is a more general sufficient condition: Any matrix that commutes with a matrix with non-degenerate eigenvalues, is diagonalizable [18].

If for one eigenvalue α of A the algebraic and geometric degeneracies are equal (say to m), then

$$A = V \begin{pmatrix} \alpha I_{m\times m} & 0 \\ 0 & A' \end{pmatrix} V^{-1}, \tag{A4}$$

where $I_{m\times m}$ is the $m\times m$ unit matrix.

An alternative representation for the matrix A is given by the singular value decomposition. Note that if det $A\neq 0$, the matrix $A[A^\dagger A]^{-1/2}$ is unitary. Then it holds

$$A=U\,[A^\dagger A]^{1/2}, \tag{A5}$$

where U is unitary. Eq. (A5) holds also for det $A=0$ via the continuity. Going to the eigen-resolution of the hermitian matrix $A^\dagger A$, we see that for *any* matrix A there is a singular value decomposition:

$$A=\sum_k \sigma_k |u_k\rangle\langle v_k|, \tag{A6}$$

$$A|v_k\rangle=\sigma_k|u_k\rangle, \qquad \langle v_k|v_n\rangle=\delta_{kn} \tag{A7}$$

$$\langle u_k|A=\sigma_k\langle v_k|, \qquad \langle u_k|u_n\rangle=\delta_{kn}, \tag{A8}$$

where σ_k (singular values of A) is the common eigenvalue spectrum of $\sqrt{AA^\dagger}$ and $\sqrt{A^\dagger A}$.

For a given diagonalizable matrix A, its singular value decomposition is related to the eigen-resolution via [18]

$$\langle v_n|R_k\rangle\sigma_n=\alpha_k\langle u_n|R_k\rangle, \tag{A9}$$

$$\langle u_n|L_k\rangle\sigma_n=\alpha_k^*\langle v_n|L_k\rangle. \tag{A10}$$

The matrix A is normal if and only if $|\alpha_k|=\sigma_k$. (I did not find any standard reference on the fact that $|\alpha_k|=\sigma_k$ leads to normality; the proof I got myself is too tedious to be presented here).

Singular values and eigenvalues are related via the Weyl inequalities. For a given matrix A, order the absolute values of its eigenvalues as $l_0\geq l_1\geq ...\geq l_n$, and order its singular values as $\sigma_0\geq\sigma_1\geq ...\geq\sigma_n$. The Weyl inequalities then read:

$$\prod_{k=0}^m \sigma_k\geq\prod_{k=0}^m l_k, \quad \prod_{k=0}^m \sigma_{n-k}\leq\prod_{k=0}^m l_{n-k}, \tag{A11}$$

$$\sum_{k=0}^m \sigma_i^\rho\geq\sum_{k=0}^m l_i^\rho, \qquad \rho>0. \tag{A12}$$

For $n=m$, (A11) leads to equality: $\prod_{k=0}^n\sigma_k=\prod_{k=0}^n l_k$.

APPENDIX B: ADDITIONAL FEATURES OF THE ENTROPY.

Recall the definitions (17) and (16) of the entropy h and the block entropy $H(N)=H(\mathcal{X}_N,...,\mathcal{X}_1)$, respectively, for the stationary process $\mathcal{X}$. Define:

$$h(N)=H(N)-H(N-1)=H(\mathcal{X}_N|\mathcal{X}_{N-1},...,\mathcal{X}_1). \tag{B1}$$

$h(N)$ [sometimes called innovation entropy] is the uncertainty of $\mathcal{X}_N$ given its history $\mathcal{X}_{N-1},...,\mathcal{X}_1$. It is clear that once $\lim_{N\to\infty}\frac{H(N)}{N}$ exists, $h(N)$ converges to the source entropy for $N\to\infty$. One can show that [6]

$$\frac{H(N)}{N}\geq h(N)\geq h(N+1)\geq h. \tag{B2}$$

To derive the second inequality in (B2) note that the stationarity and the entropy reduction due to conditioning imply

$$h(N)=H(\mathcal{X}_N|\mathcal{X}_{N-1},...,\mathcal{X}_1)=H(\mathcal{X}_{N+1}|\mathcal{X}_N,...,\mathcal{X}_2)\geq H(\mathcal{X}_{N+1}|\mathcal{X}_N,...,\mathcal{X}_1)=h(N+1). \tag{B3}$$

The first inequality in (B2) is shown as follows.

$$\frac{H(N)}{N}=\frac{1}{N}H(\mathcal{X}_1)+\frac{1}{N}\sum_{i=2}^N H(\mathcal{X}_i|\mathcal{X}_{i-1},...,\mathcal{X}_1)\geq\frac{1}{N}\sum_{i=1}^N H(\mathcal{X}_N|\mathcal{X}_{N-1},...,\mathcal{X}_1)=h(N), \tag{B4}$$

where the first equality is the obvious chain rule for the conditional information, while the second inequality in (B4) follows from the stationarity $H(\mathcal{X}_1)=H(\mathcal{X}_N)$, and then from the same reasoning as in (B3). The last inequality in (B2) is now obvious.

The meaning of $\frac{H(N)}{N} \geq h \equiv \lim_{N\to\infty} \frac{H(N)}{N}$ is that taking into account all the correlations decreases the entropy. In a related context, $h(N) \geq h(N-1)$ means that the innovations decrease under accumulation of experience. This inequality can be employed for putting an upper bound for $H(N+1)$ in terms of $H(N)$ and $H(N-1)$:

$$2H(N) - H(N-1) \geq H(N+1) \geq H(N). \tag{B5}$$

Note also that $H(N+1) = H(N) + h(N+1) \leq H(N) + \frac{H(N+1)}{N+1}$ leads to

$$\frac{H(N+1)}{N+1} \leq \frac{H(N)}{N}, \tag{B6}$$

i.e., the uncertainty per step decreases when increasing N.

APPENDIX C: ERGODIC FEATURES OF THE SINGULAR VALUES FOR A RANDOM MATRIX PRODUCT.

Let us recall some important features of the Lyapunov exponents of the random matrix product (8). Employ the known relation between the singular values of AB versus those of A and B [18]

$$\prod_{k=0}^{m} \sigma_k[AB] \leq \prod_{k=0}^{m} \sigma_k[A]\sigma_k[B], \tag{C1}$$

where $0 \leq m \leq L-1$, and where the ordering (15) is assumed: $\sigma_0[A] \geq \sigma_1[A] \geq \ldots$.

Now recall definitions (9, 10). Applying (C1) with $m = 0$ to $\mathbb{T}(\mathbf{x}_{N\ldots1})$ we get ($M < N$)

$$\ln \sigma_0[\mathbb{T}(\mathbf{x}_{N\ldots1})] \leq \ln \sigma_0[\mathbb{T}(\mathbf{x}_{M-1\ldots1})] + \ln \sigma_0[\mathbb{T}(\mathbf{x}_{N\ldots M})]. \tag{C2}$$

Thus, $\ln \sigma_0[\mathbb{T}(\mathbf{x}_{N\ldots1})]$ is sub-additive. Together with the assumptions *i)*, *ii)* and *iii)* of section IV A, Eq. (C2) ensures the applicability of the sub-additive ergodic theorem [19, 20]. This leads (for $N \to \infty$) to the probability-one convergence (24):

$$-\frac{1}{N} \ln \sigma_k[\mathbb{T}(\mathbf{x}_{N\ldots1})] \to \mu_k, \tag{C3}$$

for $k = 0$. Applying in the same way (C1) with $m = 1$ to $\mathbb{T}(\mathbf{x}_{N\ldots1})$, we use the sub-additivity for $\ln(\sigma_0[\mathbb{T}(\mathbf{x}_{N\ldots1})]\sigma_1[\mathbb{T}(\mathbf{x}_{N\ldots1})])$, deduce (24) for $k = 1$, and so on. It is clear that we could not employ the sub-additivity directly for $l_k[\mathbb{T}(\mathbf{x}_{N\ldots1})]$ (modules of the eigenvalues), since they in general do not satisfy to anything like (C1).

The sub-additive ergodic theorem is related to the additive (Birkhoff-Khinchin) ergodic theorem that claims the existence (with probability one) of a similar limit for a function $\frac{1}{N}\sum_{k=1}^{N} f[\mathcal{X}_k]$ of the stationary random process $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_N, \ldots\}$ [20].

APPENDIX D: EIGENVALUES AND SINGULAR VALUES OF THE RANDOM MATRIX PRODUCT.

Recall section IV B and the main question posed there: when the modules of the eigenvalues of the matrix product $\mathbb{T}(\mathbf{x}_{N\ldots1})$ are equal, for $N \gg 1$, to the singular values of $\mathbb{T}(\mathbf{x}_{N\ldots1})$.

As shown by (25), for $N \gg 1$ we can keep the dependence on N only in the singular values of $\mathbb{T}$. (We simplified notations as $\mathbb{T}(\mathbf{x}_{N\ldots1}) = \mathbb{T}$.) First assume that $\mathbb{T}$ is a 2×2 matrix. Write the singular value decomposition (A5) for $\mathbb{T}$ as

$$\mathbb{T} = \begin{pmatrix} e^{-N\mu_0} & 0 \\ 0 & e^{-N\mu_1} \end{pmatrix} U, \qquad U = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{D1}$$

where $e^{-N\mu_0}$ and $e^{-N\mu_1}$ [with $\mu_0 < \mu_1$] are the singular values of $\mathbb{T}$, and where the matrix U can be taken real, since $\mathbb{T}$ is real. Thus U is orthogonal: $ab + cd = 0$, $a^2 + c^2 = b^2 + d^2 = 1$, $ad - bc = \pm1$.

For the modules of the eigenvalues of $\mathbb{T}$ in (D1) one finds

$$l_0 = |a|e^{-N\mu_0} + \frac{|bc|}{|a|}e^{-N(\mu_1-\mu_0)} + \ldots, \qquad l_1 = \frac{1}{|a|}e^{-N\mu_1} - \frac{2|bc|}{|a|^3}\,e^{-N(2\mu_1-\mu_0)} + \ldots. \tag{D2}$$

If $|a| \neq 0$, the singular values of $\mathbb{T}$ coincide with the absolute values of its eigenvalues for $N \gg 1$ [23]: the terms $\mathcal{O}(e^{-N(\mu_1-\mu_0)})$ and $\mathcal{O}(e^{-N(2\mu_1-\mu_0)})$ are negligible and $\ln|a|$ is also neglected inside of the exponents as compared to $N\mu_0$ and $N\mu_1$.

This conclusion changes for $a = 0$ (and thus $d = 0$ since U is orthogonal). Now the modules of the eigenvalues coincide with each other and are equal to $e^{-N(\mu_1+\mu_2)/2}$ which is different from the singular values.

The next example is 3×3 matrix $\mathbb{T}$ with the determinant equal to zero:

$$\mathbb{T} = \begin{pmatrix} e^{-N\mu_0} & 0 & 0 \\ 0 & e^{-N\mu_1} & 0 \\ 0 & 0 & 0 \end{pmatrix} U, \qquad U = \begin{pmatrix} a & b & e \\ c & d & f \\ x & y & z \end{pmatrix}, \tag{D3}$$

where $e^{-N\mu_0}$ and $e^{-N\mu_1}$ [with $\mu_0 < \mu_1$] are two non-zero singular values of $\mathbb{T}$, and where the matrix U is orthogonal. Note that provided the third Lyapunov exponent μ_2 is larger than μ_1 (and provided we do not use the orthogonality features of the matrix U in (D3)), the considered example is sufficiently general.

Since $\det\mathbb{T} = 0$, the third singular value of $\mathbb{T}$ is zero. The third eigenvalue of $\mathbb{T}(\mathbf{x}_{N\ldots 1})$ is also equal to zero, while for the absolute values of the remaining eigenvalues we have from (D3)

$$l_0 = |a|e^{-N\mu_0} + \mathcal{O}(\frac{1}{|a|}\,e^{-N(\mu_1-\mu_0)}), \qquad l_1 = \frac{|ad-bc|}{|a|}e^{-N\mu_1} + \mathcal{O}(e^{-N(2\mu_1-\mu_0)}). \tag{D4}$$

If $|ad - bc| \neq 0$, the singular values $e^{-N\mu_0}$ and $e^{-N\mu_1}$ coincide [for $N \gg 1$] with the modules of the eigenvalues. For $|ad - bc| = 0$ the second eigenvalue of $\mathbb{T}$ is equal to zero, while the second singular value is non-zero. However, the first Lyapunov exponent is still equal to the spectral radius (module of the first eigenvalue) if $a \neq 0$. The latter two quantities are not equal for $a = 0$. Now the modules of both eigenvalues of $\mathbb{T}(\mathbf{x}_{N\ldots 1})$ reduce to $\sqrt{|bc|}\,e^{-N(\mu_1+\mu_2)/2}$.

Using the examples (D1, D3) we got a sufficient condition for deciding whether the maximal singular value of $\mathbb{T}$ is equal to the module of the corresponding eigenvalue. It is that the absolute values of the two leading eigenvalues of $\mathbb{T}$ are different.

APPENDIX E: ZETA-FUNCTION AND PERIODIC ORBIT EXPANSION.

1. Structure of periodic orbits.

Define formally

$$Z_m = \sum_{i_1,\ldots,i_m=1}^{M} \phi[A_{i_1}\ldots A_{i_m}], \tag{E1}$$

where $A_1, \ldots, A_M$ are matrices, and where $\phi[.]$ is a function that turn its matrix argument to a number. We assume that the following features hold for ϕ (d is a positive integer):

$$\phi[A^d] = \phi^d[A], \qquad \phi[AB] = \phi[BA]. \tag{E2}$$

Using these features one can prove for Z_m the following formula [26]:

$$Z_m = \sum_{n|m}\ \sum_{(\gamma_1,\ldots,\gamma_n)\in\mathrm{Per}(n)} n\,[\,\phi[A_{\gamma_1}\ldots A_{\gamma_n}]\,]^{\frac{m}{n}}\ , \tag{E3}$$

where $\sum_{n|m}$ means that the summation goes over all n that divide m, e.g., $n = 1, 2, 4$ for $m = 4$. Here $\mathrm{Per}(n)$ contains sequences

$$\Gamma = (\gamma_1, \ldots, \gamma_n) \tag{E4}$$

TABLE IV: The elements of Per(p) for $p = 1, ..., 5$ and $M = 2$. As compared to (9) we denoted $T(x_1) = 1$ and $T(x_2) = 2$. It is seen that Per(1) contains two elements, since the cyclic permutation is trivial. Per(2) contains a single element 12, since 11 and 22 remain invariant under a single cyclic permutation, while BA is obtained from AB via a single cyclic permutation. Besides the obvious sequences 1111 and 2222, Per(4) does not include the sequences 1212 and 2121 which stay invariant after two successive cyclic permutations. In Per(5) we first meet different elements that have the same overall number of 1's and 2's, e.g., 12121 and 11122.

p	Per(p)
1	1, 2
2	12
3	122, 211
4	1222, 2111, 1122
5	12222, 21111, 11222, 22111, 12121, 21212
6	122222, 112222, 111222, 111122, 111112, 112212, 221121 111212, 222121.

TABLE V: The elements of Per(p) for $p = 1, ..., 4$ and $M = 3$.

p	Per(p)
1	$1, 2, 3$
2	12, 13, 23
3	122, 211, 233 322, 133, 311 123, 132
4	1222, 2111, 1122, 2333, 3222, 2233 1333, 3111, 1133 1123, 1132, 1213 2213, 2231, 2321 3312, 3321, 3231

selected according to the following rules: *i)* Γ turns to itself after n successive cyclic permutations, but does not turn to itself after any smaller (than n) number of successive cyclic permutations; *ii)* if Γ is in Per(n), then Per(n) contains none of those $n-1$ sequences obtained from Γ under $n-1$ successive cyclic permutations.

Assume that $M = 2$, which means that the matrices A_i can take two values $A_1 = 1$ and $A_2 = 2$. With examples of Per(n) given in Table IV, the proof of (E3) is straightforward.

2. The inverse zeta-function and derivation of Eq. (44).

The inverse zeta function is defined as $\xi(z) = \exp\left[-\sum_{m=1}^{\infty} \frac{z^m}{m} Z_m\right]$, where Z_m is given by (E1). Employing (E3) and introducing notations $p = n$, $q = \frac{m}{n}$, we transform $\xi(z)$ as

$$\xi(z) = \exp\left[-\sum_{p=1}^{\infty} \sum_{\Gamma\in\mathrm{Per}(p)} \sum_{q=1}^{\infty} \frac{z^{pq}}{q} \left(\phi[A_{\gamma_1}...A_{\gamma_p}]\right)^q\right]. \tag{E5}$$

the summation over q in (E5) is taken as

$$\sum_{q=1}^{\infty} \frac{z^{pq}}{q} \left(\phi[A_{\gamma_1}...A_{\gamma_p}]\right)^q = -\ln\left[1 - z^p \phi[A_{\gamma_1}...A_{\gamma_p}]\right]. \tag{E6}$$

We shall then finally get [25, 26]:

$$\xi(z) = \prod_{p=1}^{\infty} \prod_{\Gamma\in \mathrm{Per}(p)} \left[1 - z^p \phi[A_{\gamma_1}...A_{\gamma_p}]\right] . \tag{E7}$$

3. How to generate the elements of Per(p) via Mathematica 5.

The elements of Per(p) presented in Tables IV and V were generated by hands. For larger p it is more convenient to generate these elements via Mathematica 5. Below we assume that the reader knows Mathematica at some average level. First one should run the package of combinatoric functions:

```
<<DiscreteMath‘Combinatorica‘
```

(E8)

Next one defines the function `ListNecklaces2[c_List, n_Integer?Positive]` [33], the first argument of which is a list, e.g., `{ A,B }` , while the second argument is a positive integer.

```
AllCombinations[x_List, n_Integer?NonNegative]
 := Flatten[Outer[List, Sequence  Table[x, {n}]], n - 1];
ListNecklaces2[c_List, n_Integer?Positive] := Module[{},
 Return[OrbitRepresentatives[CyclicGroup[n], AllCombinations[c, n]]]];
```

(E9)

The definition of `ListNecklaces2` proceeds via an auxiliary function `AllCombinations`. All other functions in (E9) are contained in the package (E8).

Upon running `ListNecklaces2[c, p]` one gets the elements of Per(p) together with those sequences $(\gamma_1, ..., \gamma_p)$ that remain invariant under $\bar{p}$ successive cyclic permutation, where $p/\bar{p}$ is an integer. For our purposes we meed only the sequences which are invariant with respect to p cyclic permutation, and are not variant with respect to cyclic permutations with any smaller $\bar{p}$. So our next task is to get rid of those parasitic sequences, which stay invariant with respect to $\bar{p}$ cyclic permutations with $\bar{p} < p$. To this end we designed a straightforward Mathematica program that by the direct enumeration detects and eliminates the parasitic sequences [obviously, nothing special has to be done for simple numbers like $p = 3, 5, 7, 11, 13$]. The drawback of this program is that for each p in Per(p) one has to adjust the details of this program. Anyhow, we were not able to enforce Mathematica 5 to generate the elements of Per(p) directly.

Here is an example of the above scheme: `ListNecklaces2[{A,B}, 3]` generates a list of lists:

```
{ { A,A,A }, { A,A,B}, { A,B,B}, { B,B,B} }.
```

(E10)

After elimination of the parasitic sequences this results in

```
Y = { { A,A,B}, { A,B,B} },
```

(E11)

where we introduced a shorthand `Y`. Now employing the construction

```
Apply[Times, Map[ f[#] &, Apply[Dot, Y, 1] ] ] ,
```

(E12)

where `f` is an arbitrary function, one gets

```
f[A.A.B] f[A.B.B].
```

(E13)

The construction (E12) is useful when recovering the formulas for ϕ_k for large values of p.